DOCUMENT RESUME

ED 223 680 TM 820 809

AUTHOR Dunivant, Noel
TITLE The Effects of Measurement Error on Statistical Models for Analyzing Change. Final Report.
INSTITUTION New York Univ., N.Y.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE [81]
GRANT NIE-G-78-0071
NOTE 185p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC08 Plus Postage.
DESCRIPTORS Algorithms; Analysis of Covariance; *Change; *Error of Measurement; *Mathematical Models; Performance Factors; *Psychometrics; Regression (Statistics); Research Design; *Research Methodology; *Statistical Analysis; Statistical Bias
IDENTIFIERS *Change Analysis; Linear Models

ABSTRACT

The results of six major projects are discussed including a comprehensive mathematical and statistical analysis of the problems caused by errors of measurement in linear models for assessing change. In a general matrix representation of the problem, several new analytic results are proved concerning the parameters which affect bias in observed-score regression statistics. The bias in ordinary least squares estimators is expressed as a function of covariances among true scores, among the measurement errors, and sample size. The first two projects were employed to create an algorithm for assessing the potential bias due to the unreliability of measures. The algorithm was implemented as a FORTRAN program to improve the design of investigations of change and minimize potential errors of inference. A review is presented of statistical methods which have been developed in several disciplines to estimate the parameters of true change by correcting the observed-score regression estimates for unreliability. A series of Monte Carlo experiments which evaluated the performance of the methods are discussed. The advantages and general superiority of estimators proposed by Fuller are examined. The relevance of a special linear functional relation (LFR) model and models devised for estimating the parameters of LFRs are compared. (Author/CM)

ERIC COPY

ED223680

THE EFFECTS OF MEASUREMENT ERROR ON STATISTICAL MODELS FOR ANALYZING CHANGE

Final Report
Grant NIE-G-78-0071
New York University

Noel Dunivant[1]
Principal Investigator

TM 820 809

This research was supported by Grant NIE-G-78-0071 from the National Institute of Education, U.S. Department of Education. However, points of view or opinions do not represent official NIE position or policy.

[1]Now at the National Center for State Courts, 300 Newport Avenue, Williamsburg, Virginia.

CONTENTS

CONTENTS (continued)

CONTENTS (continued)

CONTENTS (continued)

CONTENTS (continued)

CONTENTS (continued)

ABSTRACT

The final report describes the accomplishments of six major projects undertaken as part of this funded research program. First, a comprehensive mathematical and statistical analysis of the problems caused by errors of measurement in linear models for assessing change is presented. Results from several disciplines are integrated, and their implications for studies of educational change discussed. Second, a general matrix representation of the problem is formulated, and several new analytic results are proved concerning the parameters which affect bias in observed-score regression statistics. We derive equations which express the bias in OLS estimators as a function of corvariances among the true scores, covariances among the measurement errors, and sample size. Third, the results of the first two projects were employed to create an algorithm for assessing the potential bias due to the unreliability of measures. The algorithm has been implemented in the form of a FORTRAN program which can be used by researchers to improve the design of investigations of change in order to minimize the likelihood of potential errors of inference. Fourth, we undertake a comprehensive review of statistical methods which have been developed in several disciplines to estimate the parameters of true change by correcting the observed-score regression estimates for unreliability. The methods are formulated in a common algebra and evaluated in terms of bias and power. Fifth, the report describes the results of a series of Monte Carlo experiments which evaluated the performance of several methods which utilize *a priori* information about the variance structure of the errors

of measurement to estimate the parameters of the true-score regressions. The advantages and general superiority of estimators proposed by Fuller and his colleagues are discussed. Sixth, a special type of model--the linear functional relation (LFR)--is discussed in terms of its relevance for the study of change. A variety of models which have been devised in psychometrics and econometrics for estimating the parameters of LFRs are compared and recommendations about the best methods to use are made. An extensive bibliography and computer programs are included as appendices.

CHAPTER I

INTRODUCTION AND EXECUTIVE SUMMARY

OVERVIEW

This report describes the results of a large-scale research program on the effects of measurement error on linear statistical models for analyzing psycho-educational change in quasi-experimental and nonexperimental studies. As components of the research program, six projects were undertaken that analyzed the bias in observed-score regression estimators and evaluated the performance of statistical models which estimate the parameters of the true-score regressions by using information about the variance structure of the errors of measurement. Each project constitutes a separate chapter in the final report. The statistical results developed in each chapter are general in that they apply with equal validity to all linear models analyses, especially multiple regression/correlation (MRC) and analysis of covariance (ANCOVA). However, the results and their implications are discussed primarily with respect to studies of psychological and educational growth. The usefulness of the findings for educational researchers is described in considerable detail. Each chapter offers specific recommendations concerning ways in which researchers can guard against making errors of inference about the determinants of change because of errors of measurement. We believe that the results of this research program, if utilized by investigators, can substantially improve the quality of studies of educational change.

In the first project (Chapter II) we present a comprehensive mathematical and statistical analysis of the problems caused by errors of measurement in linear models for assessing change. Results from several disciplines are integrated, and their implications for studies of educational change are discussed. The second project (Chapter V) provides a general matrix representation of the problem and proves several new analytic results concerning the parameters which affect bias in observed-score regression statistics. We derive equations which express the bias in ordinary least squares (OLS) estimators as a function of the covariances among the true scores, covariances among the errors of measurement, and sample size. The objective of the third project (Chapter VI) was to devise an algorithm for assessing the potential bias resulting from the unreliability of measures. The algorithm, which has been implemented in the form of a FORTRAN program, can be used by researchers to improve the design of research projects and program evaluations in order to minimize the likelihood of potential errors of inferences about the determinants of change. As part of the fourth project (Chapter III) we undertook a comprehensive review of the statistical methods which have been developed in several disciplines to estimate the parameters of true change by correcting the observed-score regression estimates for unreliability. The methods are formulated in a common algebra and evaluated in terms of bias and power. The fifth project (Chapter VII) consisted of a series of Monte Carlo experiments which were designed to evaluate the performance of several methods that utilize <u>a priori</u> information about the variance structure of the errors of measurement to estimate the parameters of the true-score regressions. The advantages and general superiority of estimators proposed by Fuller and his colleagues are discussed. In the sixth project (Chapter IV) a special type of model--the linear functional relation (LFR)--is introduced and discussed in terms of its relevance for the study of change. A variety of models which have been developed in psychometrics and econometrics for estimating the parameters of LFRs are compared and recommendations about the best methods are made. In the following sections, a summary of each project (chapter) is given.

CHAPTER II

The purpose of this chapter is to demonstrate the bias caused of errors of measurement in linear statistical models for analyzing change and to alert educational researchers to the potential errors of inference concerning the determinants of true change which can result from using unreliable measures in multiple regression/correlation and analysis of covariance. We provide a mathematical statistical analysis of the effects of measurement error on OLS estimators. The general situation considered involves pretest and posttest measurements on some attribute that is expected to change as a function of intervening experience (e.g., treatment) and background characteristics. A general linear model, which has been proposed for studying change by several authors, is described. Definitions of parameters of change and procedures for testing hypotheses about the effects of treatment and background variables are also presented. Then a simple test score model which takes the observed (manifest) score as a linear function of true and random error (latent) variables is introduced. Next we rewrite the mathematical model of change to incorporate this measurement model, thereby explicitly recognizing the fact that the variables are not perfectly reliable.

The estimators and tests based on the observed-score distributions are then evaluated in terms of how adequately they estimate the parameters of the true-score regressions or test hypotheses about effects on true change. We prove that the observed-score regression estimators are biased and inconsistent for the structural parameters, the magnitude and direction of bias being a complex function of the intercorrelations and reliabilities of the variables. It is noted that measurement can exert harmful effects not only on estimators of the regression coefficients but also on the squared multiple correlation and mean square error. Several examples are given of how large the bias and how incorrect the resulting inferences about the determinants of change can be. Following the proofs and demonstrations, the differences between the interpretations and uses of the structural (true-score) and observed-score regression weights are discussed. We then introduce the concept of the identifiability and show how it is essential to determining the estimability of the structural parameters. In the final section of the chapter, the known conditions for the linearity of the observed-score relation when the structural relation is linear are delineated. The chapter provides a great deal of evidence that researchers should be very cautious when interpreting MRC and ANOVA results based on observed scores. In many research situations the observed-score estimates will be so biased that highly inaccurate inferences concerning the effects of treatment and background characteristics on true change will be drawn.

CHAPTER III

In this chapter a variety of single-equation statistical methods that have been developed in education, psychology, sociology, and econometrics for estimating the structural parameters are reviewed. Our objective is to draw together the techniques from diverse sources, to express them in a common algebra that is synchronous with equations of Chapter II, and to analytically evaluate them with respect to the statistical criteria of bias, power, and robustness. It is hoped that more investigators will be prompted to use one of the methods as a consequence of this review. The results are intended to serve as guides for educational researchers who wish to use one of the methods but do not know how to evaluate them.

In the first section we consider the original attenuation correction formulas of Spearman and several more recent generalizations of the method to semipartial and partial correlations. Although equations for the corrected estimators are simple and straightforward, finite sampling theory for the zero-order and partial correlations corrected for attenuation has proven intractable. The methods of Porter (1967), Stroud (1972), and DeGracie and Fuller (1972) can be used in situations appropriate for one-way analysis of covariance. Of these, Porter's and DeGracie and Fuller's procedures have the more general applicability. The exactness of Stroud's method, however, strongly commends it for the two-group design. Although the DeGracie and Fuller procedure appears less powerful than Porter's, this disadvantage may be more than offset by the reduced bias and the safeguards of the procedure which protect against the correction for attenuation producing "impossible" slope estimates.

For the more general kinds of situations in which MRC and factorial ANCOVA would be appropriate, researchers may select one of the Stouffer-Lindley or Fuller methods. It seems clear that for data which can form to the usual assumptions of normality, homoscedasticity, etc., the statistical estimation and testing procedures developed by Fuller (1980), Fuller and Hidiroglou (1978), and Warren, White, and Fuller (1974) will prove superior to the Stouffer (1936) and Lindley (1947) methods. Fuller's methods preclude estimation of singular covariance matrices following corrections due to unreliability, yield significance tests which are valid for finite samples, and provide a mechanism for incorporating information about the sampling distributions of the predictor reliabilities into the standard errors of the estimators of the true-score regression coefficients. Consequently, it appears that the Stouffer-Lindley estimators are more sensitive than those of Fuller and his associates.

It is concluded that the existing methods provide several adequate estimators of the true-score regression parameters. The major remaining problems concern sampling theory for the estimators of the structural parameters. The validity of significance tests remains a significant question for all of the estimators except Fuller's. This chapter clarifies and refines these issues, and the simulation studies reported below add further insight. It is pointed out that questions involving the type of reliability estimate to use and testing the assumption of homogenity of true-score regressions constitute important problems for future research.

Several examples of the application of the correction methods illustrated the kinds of errors of inference that could have resulted from errors of measurement in previous investigations of educational change. It is hoped that the explication and evaluation of the attenuation-correction methods provided in this chapter will encourage and facilitate their use in future studies.

CHAPTER IV

The purpose of this chapter is to analyze the problem of determining if a perfect linear relation exists among two or more variables and to review some statistical methods that have been developed to estimate and test linear functional relations. By definition, a linear functional relation (LFR) exists if the true scores on two (or more) measures are perfectly correlated.[1] Although most of the statistical work on LFR has been done by econometricians, a problem has been investigated in the field of psychometrics which is formally identical to LFR. Psychometricians have developed several statistical tests of the hypothesis that two scales measure the same attribute except for differences in means, units of measurement, and standard errors of measurement (or reliabilities). When scales satisfy these conditions they are said to be equivalent or congeneric. As is demonstrated in the chapter, equivalent tests are related by a linear functional relation. The correlation between equivalent measures, i.e., between two variables that have a linear functional relation, when corrected for attenuation (unreliability) is 1.0. In this chapter the diverse theory and methods from econometrics, statistics, education, and psychometrics are collected, compared, and integrated. Several new results are derived for the errors-in-variables problem which should prove helpful in analyzing change occuring in measures which contain errors of measurement. Several ways in which LFR models can be applied in studies of change are discussed and illustrated.

The chapter explicates and compares seven statistical methods designed to determine if the true scores from two or more tests are perfectly linearly related. They fall into one of three sets depending upon the type of information or data required by the procedure. The first group contains three methods which require replicate measures of each scale, viz., Jöreskog (1971), Kristof (1973), and Lord (1973). In the second set are three methods which assume information is available about the covariance structure of the errors of measurement. While such information may be obtained from replicated data, it can come from any other independent sources. These methods, which were formulated primarily by statisticians concerned with estimating and testing linear functional relations, include the methods of Koopmans (1937) and Tintner (1945, 1946), Fuller (1980), and Jöreskog (1971). The third set of methods includes only Fuller and Hidiroglou's (1978) procedure for testing matrix singularity when independent information about the reliabilities of the variables is available. The method uses the reliabilities to adjust the covariance matrix of observed scores in much the same way that the estimates of measurement error variances are utilized by the procedures in the second group. Indeed, all seven procedures are very similar in logic, if not in mathematical detail: each uses information about the covariance structure of the observed measures and errors of measurement (from replicate measurements, error variance estimates, or reliability estimates) to estimate the parameters of the linear functional relation.

The choice of optimal method for estimating and testing a linear functional relation depends to a great degree on the complexity of the hypothesized model and type of data available. When there are replicate measures for each variable, any one of the seven procedures can be used. With simple models, Kristof's, Fuller's, or Fuller and Hidiroglou's methods should prove generally superior to the others. Jöreskog's COFAMM and LISREL models will be preferred for more complex models where the sample size is large. The methods of Kristof and Fuller can be expected to be more robust to assumption violation, especially to nonnormality. When only estimates of measurement error variances and reliabilities are available, the relative advantages and superior performance of Fuller's methods should lead to their selection. The validity of the significance tests for finite samples strongly commends his procedures. Although Fuller's procedures may be less sensitive in certain applications, their ease of computation and the availability of a computer program for making the computations commend them.

CHAPTER V

In this chapter we derive a general matrix representation which expresses the parameters of the observed-score regression as functions of the covariances among the true and error components. Explicit expressions are derived for the bias in observed-score estimators of the mean square error, squared multiple correlation, and sampling distribution of the regression coefficients. Thus, we are able to evaluate the parameters affecting bias. The kinds of data and conditions are specified which are likely to lead to incorrect inferences concerning the determinants of true change based the results of observed-score regressions. The general equations expressing the bias in observed-score regression estimators have not been presented previously and represent a significant contribution of this research. They enable educational researchers to determine *a priori* the potential for misleading inferences in planned research.

The observed-score estimator of the mean square error is always positively biased, i.e., increased in magnitude relative to the true-score parameter, by errors of measurements in the posttest, or criterion, variable. Thus, power is reduced and the probability of Type II error is increased by unreliability of the posttest. Although a general statement can not be made about the biasing effects of correlated criterion and predictor measurement errors, with the kinds of data typically encountered in studies of educational change we can expect bias to increase as the correlations increase. As the variances of the predictor errors of measurement grow in size, the bias in the mean square error also grows. This effect becomes especially pronounced as the measurement error variances approach the magnitude of the true score variances. Unequivocal statements about the degree of bias introduced by correlations among the predictor measurement errors can not be made. It depends upon the patterns of both the true score and measurement error intercorrelations. If it is assumed that all measurement errors in the dependent and independent variables are mutually uncorrelated, it will be generally true that unreliability increases the mean square error and decreases power. Since the squared ~~multiple~~ correlation is an inverse function of the variance of the regression residual, we know that the factors which positively bias the mean square error negatively bias the estimator of the true-score multiple correlation coefficient. Although the parameters effecting bias are seen to be highly complex, the overall effect of predictor measurement errors will be to increase the bias in the estimate of the coefficient of multiple determination when most error covariances are positive and relatively small in size. Thus, observed-score regression analyses of change on the average will underestimate the goodness of fit of the true-score model in most educational applications. Bias in the estimator of the sampling distribution of the vector of observed-score regression coefficients also depends upon the covariances among the true scores and among the errors of measurement. The same factors which affect the mean square error and squared multiple correlation have similar effects on estimators of the standard errors of the regression weights. The main determinants of the joint sampling distributions, however, are the patterns of the joint distributions of the true and error components. General statements about the magnitude and direction of bias can not be made. Thus, the general effects of bias on t-tests for the individual coefficients are difficult to assess. The formulas presented in this chapter do enable researchers to evaluate the potential for bias in any specific set of circumstances, however. Therein lies their value.

CHAPTER VI

The purpose of this chapter is to develop a method for investigators to easily assess the possible impact of measurement error on statistical analyses of change. Using the results of the preceeding chapters, especially those of Chapter V, an algorithm is developed which takes as input estimates of the parameter values of the structural relations among the latent variables (which the investigator thinks are close to the true values a priori) and outputs the expected values of the corresponding observed-score regression parameters for a prespecified sample size. The logic of the algorithm is explained and illustrated with a simple example of the effects of external locus of control orientation on change in science achievement.

As part of this research program, the algorithm was implemented in the form of a FORTRAN computer program, which can be easily installed in most software libraries. Input to the program consists of information about the covariances among the true predictors, the reliabilities of the observed predictors, and the true-score regression coefficients. The program outputs values of the true-score regression parameters and those of the corresponding observed-score regression parameters. Comparison of the two sets of parameter values allows one to assess the degree of bias likely to occur in observed-score regression coefficients as estimators of their true-score counterparts. In the final section of the chapter, a comprehensive application of the computer program is presented.

Use of the program will enable investigators to become aware of the ways in which measurement error may bias regression analyses of change. Making this evaluation before data collection is completely analogous to carrying out a power analysis. The results of the assessment may lead the investigator to modify data collection plans. For example, the program may reveal that the reliability of the pretest must be increased if accurate inferences are to be possible. The assessment may indicate that bias can not be avoided easily and prompt the investigator to gather the data in such a way as to make the use of attenuation-correction methods or multiple indicator (LISREL) models possible. Also, as with power analysis, the program can be used <u>post hoc</u> to determine the degree of caution one should have when interpreting the results of the regression analyses of observed scores. In many situations, like the one described in the example in this chapter, it will be concluded that the possible bias in the observed-score regression estimators was so great that any inferences must be regarded as completely suspect.

CHAPTER VII

Chapter VII reports the results of the Monte Carlo experiments designed to evaluate the performance of various multiple regression and analysis of covariance methods that correct for errors of measurement. The objective of this research was to determine which statistical procedures for estimating the structural parameters of change demonstrated the least bias and most power. Only when this information is provided to educational researchers can they choose an estimation technique that is optimal for their purposes. The results of these simulations can be utilized to reduce the chances for drawing faulty conclusions about the effects of treatments or individual differences on true change analyses of observed scores. We were fortunate to be able to derive a number of analytic results which obviated the need for some of the simulations that had been originally anticipated.

Two simulation experiments were conducted. The first compared regression methods, and the second evaluated analysis of covariance procedures. Different simulations were required because the MRC and ANCOVA correction methods required different information. Specifically, individual scores were needs for the ANCOVA simulations, while only covariance matrices were required for the regression studies.

In the first experiment the performance of the Stouffer-Lindley method was compared with that of the Warren, White, and Fuller (1974) procedure. Both of these were contrasted with traditional OLS regression analysis of the observed scores. Covariance matrices of true, error, and observed scores were defined according to the equations given in the preceding chapters and randomly generated using IMSL subroutines. The effects of six factors on the relative performances of the OLS, Stouffer-Lindley, and Warren, White, and Fuller methods were systematically assessed: the effect of the pretest, the pretest-research factor correlation, the pretest reliability, the posttest reliability, the sample size, the effect of the research factor on true change, the sampling variance of the pretest reliability coefficient. The performance of the methods was measured for many statistics, including the mean square error of the model, pretest regression coefficient, and the standard error of the pretest regression weight. Main interest, however, concerned the bias and sensitivity of the regression estimator for the effect of the research factor on change. Bias, power, and probability of Type I error for the three methods were evaluated as relative and absolute criteria. The results indicated that both the Stouffer-Lindley and Warren, White, and Fuller methods performed adequately across the conditions simulated. The degree of bias in both null and nonnull conditions was small, usually less than 10%. The direction of the bias, however, was unpredictable. In general, bias increased as the effect of the pretest (or, pretest-posttest correlation) and the pretest-research factor correlation increased. Bias decreased as the reliability of the research factor grew. Empirical alpha values did not differ substantially from nominal levels. Power was enhanced by pre-post correlation, reliability of both pretest and research factor scores, and measurement. As expected, pretest-research factor correlation adversely affected power. The Warren, White, and Fuller method was superior to the Stouffer-Lindley procedure when the sampling variability of the estimators of the pretest measurement error variance (or reliability) was recognized. The former method explicitly incorporates information concerning the variability of reliability estimators. As the sample size upon which the reliability estimate is based becomes very large, the two methods produce virtually identical results. With small sample sizes, however, the Stouffer-Lindley estimators can perform very poorly under certain sets of conditions. We conclude that the method of Warren, White, and Fuller can be recommended as the multiple regression method of choice for studies of educational change.

Results from the second series of Monte Carlo studies on ANCOVA methods produced results that closely parallel those obtained for regression methods. The effects of several factors, e.g., pretest-posttest correlation, on the bias and power of estimators of covariate-adjusted means were assessed in the two-group ANCOVA design. The DeGracie and Fuller method demonstrated superior performance to the Cohen and Cohen and Porter methods when there was variation in reliability estimate and sample sizes were small. The differences in performance among the methods diminished as sample size increased. The use of the DeGracie and Fuller ANCOVA method for estimating the effects of treatment groups on true change is advocated for the kinds of situations generally found in educational research. The availability of a computer program for performing regression and covariance analysis by the Fuller methods greatly facilitates the application of these true-score estimation procedures in studies of educational growth.

CHAPTER II

PROBLEMS CAUSED BY MEASUREMENT ERROR IN ANALYZING CHANGE

INTRODUCTION

The objective of this chapter is to provide a mathematical analysis of the effects of measurement error on statistical models for analyzing change. The general situation that we consider involves pretest and posttest measurements on some attribute that is expected to change as a function of intervening experience (e.g., treatment) and background characteristics. A general linear model which has been proposed by several authors for studying change is presented, and definitions of parameters of change and procedures for testing hypotheses about change as a function of treatment and background characteristics are developed. Then a simple test score model which takes the observed score as a linear function of an unobserved true (or latent) variable and a random error component is introduced. The mathematical model of change is then rewritten to incorporate this measurement model, thus explicitly recognizing the fact that the variables are not perfectly reliable.

The estimators and tests based on the observed-score distributions are then evaluated in terms of how adequately they estimate the parameters of the true-score distribution or test hypotheses about true change. Briefly, it is proved that the observed-score estimators are biased for the structural parameters, the magnitude and direction of bias being a complex function of the intercorrelations and reliabilities of the variables. Next covariance structure analysis is used to explicate the relationship between the observed-score and latent-variable parameters. This enables us to determine values of the observed-score parameters when given the corresponding values of the true-score parameters and reliability information.

It should be pointed out that in this chapter we do not treat those situations where assignment to treatment has been made on the basis of an unreliable pretest. Following the pioneering work of Goldberger (1972), several statisticians (Kenny, 1975; Overall & Woodward, 1976a; Rubin, 1977; Weisberg, 1979) have demonstrated that unbiased estimators of the differences between the treatment and control groups in true change can be obtained when this kind of selection process is employed. Furthermore, the effects of measurement error when group regressions have heterogeneous slopes lies beyond the scope of this chapter. The reader is referred to Rogosa (1977b) for an excellent, comprehensive treatment of the effects of measurement error on the Johnson-Neyman technique.

GENERAL STRUCTURAL MODEL OF CHANGE

Before discussing the bias caused by errors of measurement the general framework and model for studying change must be developed since this has been an issue of some controversy (cf. Cronbach & Furby, 1970; Keesling & Wiley, 1977; Linn & Slinde, 1977; Werts & Linn, 1970; Wiley & Harnischfeger, 1973). Following Wiley and Harnischfeger (1973), we wish to consider a structural model of the effects of initial status (pretest), treatment program, and background characteristics on final status (posttest) with respect to some quantitative attribute.

The model diagrammed in Figure 1 shows the possible causal relations (assuming the model is correctly specified) among these four factors.

Insert Figure 1 about here

Since the model is represented along a time dimension, the background variables, e.g, parents' SES, sex, or aptitude, are depicted as most remote from the posttest. These background characteristics can affect either initial status (solid arrow) or the probability of assignment to treatment programs when subjects are not randomly assigned (broken arrow). Initial status can influence final status (solid arrow) and treatment groups' membership when assignment to groups is based on pretest scores (broken arrow).

Letting X_1 represent initial status, X_2 treatments, X_3 background characteristics, and Y final status, the model may be written in symbolic form as:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e \quad (1)$$

where b_j are structural coefficients characterizing the multivariate distribution of the Y and X_j, and e is a stochastic term symbolizing sampling or specification error. In most applications the b_j are partial regression weights indicating the contribution of X_j to Y and could be subscripted as $b_{YX_j.X_{j'}}$. This notation more clearly demonstrates that reference is to the effect of X_j on Y while controlling the effects of the other $X_{j'}$ (where $j' \neq j$).

If we define change or growth as the difference between final and initial statuses, simple algebraic manipulation of Equation 1 allows us to show the relationship of this model to one that takes change as the dependent variable. Subtracting X_1 from both sides of Equation 1, we obtain:

$$C = Y-X_1 = b_0 + (b_1-1)X_1 + b_2X_2 + b_3X_3 + e \quad (2)$$

where C designates change. Thus, b_2 and b_3 are the same in both equations while the weight for the pretest in Equation 2 is simply one unit less than the comparable coefficient in Equation 1. With this approach to defining change, clearly there is no need to deal with actual change scores as Werts and Linn (1970) and Wiley and Harnischfeger (1973) have pointed out. The coefficients specified in Equation 1 can be interpreted as parameters of change.

The growth model represented in Equations 1 and 2 is defined in terms of true scores or latent variables. That is, the equations specify the structural model in the same terms as does the theory, i.e., as relations among hypothetical constructs (cf. Cronbach & Meehl, 1955). In this context, the b_j assume considerable importance as parameters of the hypothetical mechanism which generates the observed data. Thus, the b_j indicate the strengths of particular connections among theoretical constructs and, collectively, define a behavioral or psychological law. In almost all research on change the investigator seeks information about the form of the structural model defining change in status for some behavioral domain and estimates the magnitudes of the b_j. If the constructs or latent variables could be precisely measured, it would be a relatively simple procedure to estimate the b_j from the measurements and to evaluate the adequacy of the model in accounting for the observed data. Unfortunately, in the social sciences our capability to measure theoretical constructs without error is limited so that estimating and testing structural models becomes highly problematic.

To describe the mechanism which more adequately reflects our beliefs about how the observations came into being, we construct the model depicted in

Insert Figure 2 about here

Figure 2. In this schematic, recognition is given the fact that one's measurements are fallible, i.e., contain errors of measurement. Now a system of structural equations is required to specify the model.[1] We add to Equation 1 the following:

$$x_1 = f_{10} + f_{11}X_1 + u_1 \tag{3}$$

$$x_2 = f_{20} + f_{21}X_2 + u_2 \tag{4}$$

$$x_3 = f_{30} + f_{31}X_3 + u_3 \tag{5}$$

$$y = f_{y0} + f_{y1}Y + v \tag{6}$$

where x_j and y are observed or measured values, the X_j and Y are true scores or latent variables, the u_j and v are errors of measurement, and the f coefficients are parameters specifying the regression of the observed scores on the underlying factors. The f_{ij} are, in fact, factor loadings. The reader may recognize this as an application of Jöreskog's (1971) theory of congeneric tests. Using vector and matrix notation Equations 1 through 6 may be written compactly as

$$Y_k = \underline{b}\,^t\underline{X}_k + e_k \tag{7}$$

$$\underline{x}_k = \underline{f}_{0x} + \underline{F}_x X + \underline{u}_k \tag{8}$$

$$y_k = \underline{f}_{0y} + \underline{F}_y Y + \underline{v}_k \tag{9}$$

where underscoring is used to designate vector (lower case) and matrix (upper case) quantities[2] and $\underline{F}_x$ is a diagonal matrix with the factor loadings on the principal diagnoal. Together Equations 7, 8 and 9 constitute a LInear

Structural RELations (LISREL) model as defined by Jöreskog (1972, 1973). Equations 8 and 9 specify what is termed the measurement model, while Equation 7 represents the structural or causal relation. Although social researchers state their hypotheses in terms of Equations 7-9 and would like to estimate the values of b_j and f_{ij} contained therein, most are forced to perform a regression analysis of the observed scores (treating them as if they were the true scores). This situation is depicted in Figure 3, which shows that the x_j have been substituted for the X_j. The regression parameters giving the

Insert Figure 3 about here

expectation of y for fixed x_j are designated with primes, b'_j, to indicate their correspondence to the respective structural parameters, b_j. Patently, the b'_j will equal the corresponding b_j only under a very limited set of conditions. Using estimators of the b'_j as estimators of the b_j is unsatisfactory in most applications, because the $\hat{b}'_j$ are neither unbiased nor consistent estimators of the b_j. Therefore, inferences about the nature of change and its determinants can be inaccurate or misleading if based on the regression estimators, $\hat{b}'_j$. In the next section we demonstrate the bias and inconsistency of the observed-score estimators and describe the potentially deleterious effects of measurement error on inferences about change.

PROOF OF BIAS AND INCONSISTENCY

In this section we consider the consequences of using the $\hat{b}'_j$ as estimators of the structural coefficients. For ease of exposition, the case of one X variable is taken up first and then the two-predictor case. Following proofs for these cases, the proof of bias and inconsistency for the general multiple variable situation is derived.

The Single-Predictor Case

To prove that the Ordinary Least Squares (OLS) regression analysis of the observed scores produces biased and inconsistent estimators of the structural parameters when there is a posttest and a single pretest, we begin by writing:

$$y = b'_0 + b'_1 x_1 + e' , \tag{10}$$

and letting $f_{01} = 0$ and $f_{11} = 1.0$ in Equation 3 and $f_{y1} = 0$ and $f_{y1} = 1.0$ in Equation 4

$$x_1 = X_1 + u_1 , \tag{11}$$

$$y = Y + v . \tag{12}$$

$$\hat{b}'_1 = \frac{(x_{1k} - \bar{x}_1)(y_k - \bar{y})}{\Sigma_k (x_{1k} - \bar{x}_1)^2} = \frac{\hat{s}_{yx_1}}{\hat{s}_{x_1}{}^2} , \tag{13}$$

and $$\hat{b}'_0 = \bar{y} - \hat{b}_1 \bar{x}_1 . \tag{14}$$

The summation is taken over k = 1 to N units of observation. Substituting Equations 11 and 12 into Equation 13, we find

$$\hat{b}'_1 = \frac{\Sigma[(Y + v) - (\bar{Y} + \bar{v})]\,[(X_1 + u_1) - (\bar{X} + \bar{u}_1)]}{\Sigma[(X_1 + u_1) - (\bar{X}_1 + \bar{u}_1)]^2}$$

$$= \frac{\Sigma(Y-\bar{Y})(X_1-\bar{X}_1) + \Sigma(X_1-\bar{X}_1)(v-\bar{v}) + \Sigma(Y-\bar{Y})(u_1-\bar{u}_1) + \Sigma(v-\bar{v})(u_1-\bar{u}_1)}{\Sigma(X_1-\bar{X}_1)^2 + 2\Sigma(X_1-\bar{X}_1)(u_1-\bar{u}_1) + \Sigma(u_1-\bar{u}_1)^2} \tag{15}$$

On the assumption that $E(X_1 u_1) = E(X_1 v) = E(Y u_1) = 0$, the second and third terms in the numerator and the second term in the denominator approach zero as the sample increases without limit. Thus, the probability limit for $\hat{b}'_1$ is given as

$$\text{plim } \hat{b}'_1 = \frac{s_{YX_1} + s_{vu_1}}{s_{X_1}^2 + s_{u_1}^2} \quad . \tag{16}$$

Making the additional assumption that errors of measurement in y and x_1 are uncorrelated and noting that by definition we divide both the numerator and denominator of Equation 16 by s_X^2 to obtain

$$b_1 = \frac{s_{YX}}{s_X^2} \tag{17}$$

$$\text{plim } \hat{b}'_1 = \frac{b_1}{1 + \dfrac{s_{u_1}^2}{s_{X_1}^2}} \quad . \tag{18}$$

The probability limit of $\hat{b}'_1$ does not equal b_1 but underestimates it. Thus, the OLS estimator of the slope of the regression of the posttest on pretest is an <u>inconsistent</u> estimator of the structural parameter. By similar steps and noting that $E(e - \hat{b}_1 u_1)(x_1 - \bar{x}_1)$ does not equal zero, it follows that $\hat{b}'_1$ is a <u>biased</u> estimator of b_1.[3] From psychometric theory (Lord & Novick, 1968) the population reliability of the observed variable x_1 can be written as

$$r_{11} = \frac{s_{X_1}^2}{s_{X_1}^2 + s_{u_1}^2} = \frac{s_{X_1}^2}{s_{x_1}^2} \quad . \tag{19}$$

Using this identity Equation 18 can be rearranged into a more familiar form:

$$\text{plim } \hat{b}'_1 = \frac{b_1}{\dfrac{s_{x_1}^2 + s_{u_1}^2}{s_{X_1}^2}} = \frac{\dfrac{s_{X_2}^2}{s_{X_1}^2 + s_{u_1}^2}}{1} b_1 = r_{11} b_1 \quad . \tag{20}$$

Alternatively,

$$b_1 = \frac{b'_1}{r_{11}} \qquad (21)$$

For additional details concerning this proof the interested reader is referred to Bohrnstedt (1969, pp. 122-125), Cochran (1968, pp. 651-652), Johnston (1963, pp. 148-150), and Schmidt (1976, pp. 105-115).

Since r_{11} 1.0, b'_1 b_1. The relationship (structural coefficient) between the latent variables is always greater on average than which would be inferred from the OLS regression of observed scores when these observations are fallible. Thus, errors of measurement "attenuate" the regression of Y on X_1, that is, they bias the estimate of the slope toward zero. Note that it is the errors in x and not those in y which cause the bias as long as $E(vu_1) = 0$. Since a value of 1.0 for $\hat{b}_1$ would indicate no expected change from Time 1 to Time 2 measurements, the attenuated estimator $\hat{b}'_1$ will lead to the faulty inference that persons above and below the pretest mean ($\bar{X}_1 = \bar{x}_1$) will show more absolute change than is actually the case. Of course, this is the well-known regression to the mean phenomenon caused by errors of measurement (Campbell & Stanley, 1963). The point is that inferences about true change and about true change as a function of true initial status (Thomson, 1924; Werts & Hilton, 1977) will be inaccurate because of the unreliability of the pretest measurements.

The Two-Predictor Case

It is now our purpose to demonstrate the bias and inconsistency in the $\hat{b}'_j$ two predictor case where X_1 is a pretest and X_2 represents another determinant of change, either a treatment or background variable. If this second variable is classificatory, i.e., represents membership in a treatment or sociodemographic group, then X_2 becomes a coded variable, and the b_2 is a function of the mean differences between groups. Thus, analysis of covariance can be represented as a standard multiple regression problem (Cohen & Cohen, 1975; Overall & Klett, 1972). Our concern is with the structural model specifying change as a function of initial status and treatments or background characteristics. This is most easily dealt with by expressing Y as a function of X_1 and X_2:

$$Y = b_0 + b_1X_1 + b_2X_2 + e \quad . \tag{22}$$

We begin with the standard definitions of the structural parameters as given in any advanced text on linear models:

$$b_1 = \frac{s_{X_2}{}^2 \; s_{YX_1} - s_{X_1X_2} \; s_{YX_2}}{s_{X_1}{}^2 \; s_{X_2}{}^2 - s_{X_1X_2}{}^2} \quad , \tag{23}$$

$$b_2 = \frac{s_{X_1}{}^2 \; s_{YX_2} - s_{X_1X_2} \; s_{YX_1}}{s_{X_1}{}^2 \; s_{X_2}{}^2 - s_{X_1X_2}{}^2} \quad , \tag{24}$$

and

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 \quad . \tag{25}$$

Adding the stipulations that $x_2 = X_2 + u_2$ and that $E(u_1u_2) = E(u_2v) = E(X_1u_2) = E(Y_2) = E(Y_1) = 0$ to the bivariate regression model considered above, we can derive the expressions for the expected values of b'_j as functions of the b_j. The formulas for the observed-score estimators analogous to Equations 23-25 are

$$\hat{b}'_1 = \frac{\hat{s}_{x_2}{}^2 \; \hat{s}_{yx_1} - \hat{s}_{x_1x_2} \; \hat{s}_{yx_2}}{\hat{s}_{x_1}{}^2 \; \hat{s}_{x_2}{}^2 - \hat{s}_{x_1x_2}{}^2} \quad , \tag{26}$$

$$\hat{b}'_2 = \frac{\hat{s}_{x_1}^2\ \hat{s}_{yx_2} - \hat{s}_{x_1x_2}\ \hat{s}_{yx_1}}{\hat{s}_{x_1}^2\ \hat{s}_{x_2}^2 - \hat{s}_{x_1x_2}^2}, \tag{27}$$

and

$$\hat{b}'_0 = \bar{y} - b'_1\bar{x}_1 - b'_2\bar{x}_2 \ . \tag{28}$$

Using the psychometric identity that $s_{YX_j} = s_{yx_j}$ under the stated assumptions (Lord & Novick, 1968) and substituting expressions for the true scores into Equations 26 and 27 the following results are obtained:

$$\hat{b}'_1 = \frac{\widehat{(s_{X_2}^2 + s_{u_1}^2)}\ \hat{s}_{YX_1} - \hat{s}_{X_1X_2}\ \hat{s}_{YX_2}}{\widehat{(s_{X_1}^2 + s_{u_1}^2)}\ \widehat{(s_{X_2}^2 + s_{u_2}^2)} - \hat{s}_{X_1X_2}^2}, \tag{29}$$

and

$$\hat{b}_2 = \frac{\widehat{(s_{X_1}^2 + s_{u_1}^2)}\ \hat{s}_{YX_1} - \hat{s}_{X_1X_2}\ \hat{s}_{YX_1}}{\widehat{(s_{X_1}^2 + s_{u_1}^2)}\ \widehat{(s_{X_2}^2 + s_{u_2}^2)} - \hat{s}_{X_1X_2}^2} \ . \tag{30}$$

Clearly the b'_j do not approach the b_j in the limit unless $s_{u_1}^2 = s_{u_2}^2 = 0$. Thus, the b'_j are not consistent estimators for the structural coefficients. As with the bivariate case, the bias of the b'_j follows from the fact that the expected value of the covariance of the residuals from regression and the true X_j values does not equal zero. Equations 29 and 30 reveal how potentially misleading the observed-score regression weights can be as estimators of the structural coefficients. The value of b'_1 can be greater or less than b_1 depending upon the magnitude of $s_{u_1}^2$. A similar result holds for b'_2 and b_2. While in the bivariate regression case the slope estimator is attenuated by errors of measurement on the average, in the multiple regression case the value of b'_j can be either attenuated or "accentuated" (Wiley & Hornik, 1972) by measurement or observational errors.

The fact that b'_j can be statistically significant with a negative sign while the structural parameter has a relatively large positive value has lead many statisticians and psychometricians to recommend strongly against the use of b'_j as estimators of the b_j, e.g., Cochran (1968), Cohen & Cohen (1975), Cronbach (1976), Hummel-Rossi & Weinberg (1975), and Lord (1958). The bias in the observed-score regression weights attributable to measurement error certainly poses a grave problem for longitudinal research. It can produce inferences that the effect of Treatment 1 relative to Treatment 2 was harmful when in truth it was beneficial. Or, it may lead to conclusions like high SES children changed more than low SES children when, in fact, relative change was in the opposite direction. This would seem an intolerable state of affairs. Moreover, the biases in the regression weights are not the only parameters of change affected by measurement error. These will be described after the expression of the bias in the b'_j has been derived for the general case of J predictors.

The General Case

Let $\underline{S}_{XX}$ and $\underline{S}_{xx}$ be the variance-covariance matrices of the true and observed scores, $\underline{S}_{uu}$ the covariance matrix of the errors of measurement in the x variables, $s_{XY} = s_{yx}$ the vectors of predictor-criterion covariances and $\underline{b}$ and $\underline{b}'$ the vectors of regression weights for the true and fallible variables, respectively. Then it follows that

$$\underline{b}\,\underline{S}_{XX} = \underline{s}_{YX} \quad , \tag{31}$$

$$\underline{b}'\,\underline{S}_{xx} = \underline{s}_{yx} = \underline{s}_{YX} \quad , \tag{32}$$

$$\underline{b}' (\underline{S}_{XX} + \underline{S}_{uu}) = \underline{s}_{YX} \quad , \tag{33}$$

and
$$\underline{b}\ \underline{S}_{XX} = \underline{b}'(\underline{S}_{XX} + \underline{S}_{uu}) \quad , \tag{34}$$

where the vectors $\underline{b}$ and $\underline{b}'$ are transposed to make the quantities conformable for multiplication. Postmultiplying both sides of Equation 34 by $\underline{S}_{XX}^{-1}$, we find that $\underline{b}$ can be written as a weighted function of $\underline{b}'$ (cf. Lindley, 1947, pp. 227, Eq. 40):

$$\underline{b} = \underline{b}'\ \underline{S}_{xx}\ \underline{S}_{XX}^{-1} \quad . \tag{35}$$

This matrix product indicates that unless the X_j are uncorrelated, the bias in b'_j as an estimator of b_j depends not only on the errors in x_j but also on all intercorrelations $X_{jj'}(j \neq j')$.[4] Thus, even when X_j is measured perfectly so that $x_j = X_j$, b'_j still will be a biased estimator of b_j.

Errors of measurement, therefore, should exert their most damaging effects in survey and quasi-experimental studies where the lack of experimental control will result in substantial intercorrelations among the pretest and factors associated with change. In these kinds of studies of change the weight associated with the pretest "undercontrols" (Wiley & Hornik, 1972) or insufficiently adjusts for differences among individuals in initial status. In ANCOVA the regression correction for covariate (pretest) differences between groups may be too little or too great, and the resulting comparison of differences in adjusted posttest means will be biased if the groups differed initially (Snedecor & Cochran, 1967). In predictor sets where some of the variables are more reliable than others, part of the contribution of the less reliable predictor will be attributed to the more reliable predictors. However, other factors associated with quasi and nonexperimental studies may more than offset the inadequate adjustment due to measurement error (Cronbach et al., 1977; Heckman, 1979; Olejnik & Poeter, 1981; Weisberg, 1979). Under some circumstances, no adjustment at all for the pretest reduces the bias in b_j.

EFFECTS OF MEASUREMENT ERROR ON OTHER ESTIMATES

In addition to the problems with the regression weights there are other potential pitfalls to interpreting the observed-score regression results as providing verdical information about the hypothesized structural model. Some of these will be briefly summarized. First, errors of measurement will make the overall adequacy of the model appear less than if the variables were perfectly reliable. Both indices of the goodness of fit of the data to the model, the coefficient of multiple correlation or determination (R^2), and the mean square error (MSE), or residual variance, will be biased by observational errors. R^2 will be attenuated and MSE inflated on the average (Bohrnstedt, 1969; Cochran, 1968, 1970). When the X_j are uncorrelated, Cochran (1970) has shown that the degree of attenuation in R^2 is a function of the reliabilities of y and the x_j:

$$R'^2 = R^2 \, r_{yy} \, \bar{r}_w \tag{36}$$

where $\bar{r}_w$ is a weighted average of the r_{jj}. It is apparent from the formula for the mean square error or the residual variance from regression,

$$s_e{}^2 = s_y{}^2 \, (1-R'^2) \tag{37}$$

that errors of measurement will have a proportionally greater effect as R^2 increases. Although he was unable to derive a closed form expression for R'^2 when the X_j are correlated, Cochran (1970) suggests that Equation 36 "may serve as a rough guide to the effect of errors of measurement on the squared multiple correlation in many applications. The value of R'^2 may be up to 10 percent higher (than Equation 36) if most correlations are positive and harmful and the r_{jj} exceed .7, and up to 25% higher if the r_{jj} are as low as .5." The decrease in explained variation means that the power of statistical tests will be lowered. While the errors in y do not contribute to the bias in the b'_j as long as v is uncorrelated with Y, X_j and u_j, they do contribute to the reduction in R^2 as indicated in Equation 36 and thus also to loss of power (cf. Bohrnstedt, 1969; Cleary & Linn, 1969; Cochran, 1968, 1970; Nicewander & Price, 1977; Sutcliffe, 1958; Walker & Lev, 1953; Winne, 1977).

Although the raw partial regression weights are not biased by errors in y, the standardized partial regression weights and path coefficients are attenuated by v. Additionally, both prediction and simulation will be affected by errors in y.

The final consequence of errors in measurement in analyses of change concerns the distortions they cause in analysis of covariance. As pointed out previously, errors of measurement in the pretest will bias the estimates of adjusted posttest differences if the groups differ in mean pretest scores (Campbell & Erlebacher, 1970; Dunivant 1975, 1977; Kenny, 1975; Overall & Woodward, 1976a, b; Rubin, 1977; Werts & Linn, 1971). ANCOVA is predicated on the assumption of homogeneous pooled within-groups regressions of covariate on criterion. Whenever the slopes are heterogenous, or equivalently, there is a covariately-research factor interaction, ANCOVA is no longer appropriate. A mathematical model which evaluates the differences in regression lines must be adopted. This may take the form of the Johnson-Neyman technique or product-vectors in multiple regression.

Recently, Rogosa (1977a, b) has demonstrated that the loss of power due to errors of measurement will cause the investigator to fail to reject the hypothesis of homogeneity of regression in many situations where it is false (Type II error). Thus, ANCOVA will be utilized on many occasions when the Johnson-Neyman technique or analysis of partial variance (Cohen & Cohen, 1975) are appropriate. Faulty inferences about the underlying causal model will frequently result. The reader is referred to Rogosa's (1977a, b) papers for a presentation of the biasing effects of measurement error on the Johnson-Neyman technique. A recent search of the literature (see appendix) located only one reference (Busemeyer, 1980) out of over 400 articles surveyed which treated the effects of measurement error on estimators of nonadditive or interactive effects in multiple regression (Dunivant, 1980). It is fair to conclude from the demonstrations presented in this section, that in analyses of change, the potential for errors of inference caused by errors of measurement is very great. This problem should be of considerable concern to any investigator who collects test-retest data and wishes to explain change in scores during the interval. In the following we review several examples of the magnitude of the bias in b'_j as estimators of b_j.

DEMONSTRATIONS OF BIAS CAUSED BY ERRORS OF MEASUREMENT

Several statisticians have constructed hypothetical examples to illustrate the kinds of problems caused by errors of measurement. Cochran (1968) provided the coefficients reproduced as Table 1.

Inspection of the entries reveals that in the two predictor case when $b_1 = 2$ and $b_1 = 1$, and that the estimators b'_1 and b'_2 may simultaneously underestimate both b_1 and b_2 or overestimate one and underestimate the other depending upon the respective reliabilities of x_1 and x_2. There is even one example ($r_{11} = .6$, $r_{22} = 1.0$, $r_{X_1X_2} = .3$) where $\hat{b}'_1 < \hat{b}'_2$. In an example constructed by Bornstedt and Carter (1971) the observed-score estimate and structural coefficient had opposite signs. If $r_{11} = r_{22} = .81$, $r_{yy} = 1.0$, $r_{yx_2} = .7$, and $r_{yx_1 \cdot x_2} = .031$, then $b'_1 = .03$ while $b_1 = -.186$.

A different approach to demonstrating the bias in partial regression coefficients has been pursued by Corder-Bolz (1978), Hanushek and Jackson (1977), Ladd (1956), Marston and Borich (1977), McLean, Ware and McClave (1975), and Porter (1967). These statisticians have employed Monte Carlo or simulation techniques to generate data which conform to a structural model whose parameter values are specified a priori. Hundreds or thousands of samples of simulated observations are then analyzed by OLS regression and the mean and variance of the resulting b'_j are compared with the preset b_j. Thus, Hanushek and Jackson (1977) generated 100 samples of 200 observations each from a model with parameters $b_0 = 15$, $b_1 = 1$, $b_2 = 2$, and $r_{X_1X_2} = 0$. When r_{22} equaled .8, the mean estimates were $b'_1 = .99$ and $b'_2 = 1.31$. In another simulated experiment the means were .97 and .36 for b'_1 and b'_2, respectively, when the reliability of x_2 was lowered to .4.

Corder-Bolz (1978), McLean et. al. (1975), Marston and Borich (1975), and Porter (1967) investigated the effects of measurement error in the covariate on tests of adjusted group differences in ANCOVA. McLean et al.

(1975) varied the reliabilities of the covariate within and across groups, the sample size, and the mean differences on the covariate between the experimental and control groups in a 6 x 2 x 2 x 2 factorial ANCOVA design. Both the empirical alpha level (Type I error probabilities) and empirical power (1-Type II error probabilities) of the hypothesis of no adjusted posttest differences were evaluated for 2000 sets of generated observations. The results indicated that if the groups' pretest means were equal (no pretest-group factor correlation), then the nominal alpha values were not significantly disturbed by errors in the covariate as would be expected (Kenny, 1975; Overall & Woodward, 1976a, b). However, if there was a pretest-treatment correlation, then the nominal alpha values were greatly affected. In general, the fallibility of the covariate resulted in an underadjustment of posttests differences so that the empirical alphas exceeded the nominal alphas. With reliabilities in the .5 range, Type I errors were made in 40% to 100% of the samples depending upon whether the n per group was 10 or 100. For all conditions empirical power differed significantly from true power. Sometimes the empirical power was significantly lower and sometimes it was significantly greater than the theoretical value. According to McLean et. al. (1975) "the most dramatic result (was) that where the experimental group actually experienced a gain and the control group did not and the pretest mean of the experimental group was less than that of the control group, the adjusted posttest means indicated that the control group was better" (p. 550).

After conducting extensive simulations in which reliabilities, pre-posttest correlations, covariate-treatment correlations, and treatment effects were varied systematically through a wide range of values, Corder-Bolz (1978) concluded "that the models traditionally used to evaluate change [including ANCOVA] can produce seriously distorted results" (p. 975). Porter (1967) also conducted extensive simulations. His work indicated that the bias in ANCOVA estimators and significance test could be very large. In contrast to these findings which accord closely with the derivations presented above, Marston and Borich (1977) reported that in their Monte Carlo investigation of ANCOVA with an unreliable covariate, the tests of adjusted group differences did not exceed the nominal alpha level, even when the pretest means differed. It is difficult to explain this anomalous result assuming that their data generation procedure performed as they expected. Using a complete, over-identified, nondynamic two-equation model Ladd (1956) generated 30 samples of observations on two endogeneous and two exogeneous variables. The reliabilities of the variables ranged from .74 to .92. When the two regression equations were estimated separately by OLS in the 30 samples Ladd found that the average b'_j sometimes was greater than and sometimes smaller than their respective structural parameter values. The average least squares bias ranged from 0 to 31% for the eight regression coefficients across the 30 samples.

To summarize the results of the Monte Carlo demonstrations, in four of five investigations, the findings of the simulated data were congruent with the proofs and derivations presented in the previous sections. Thus, the evidence is quite substantial that if the observations are actually generated by a mechanism modeled by Equations 1-6 (or 7-9), OLS regression estimates derived from observed scores will lead to errors of inference because of the errors of measurement.

Before concluding this section it is instructive to note that several writers have illustrated the possible bias due to the "errors in variables" in terms of zero-order and partial correlations. Since these correlations can be written as functions of the b_j, the same bias will be observed in the r' as in the b'_j. Consider the one predictor case where the zero-order correlation between true pretest and true posttest can be written

$$r_{YX_1} = b_1 \frac{s_{X_1}}{s_Y} \tag{38}$$

and the corresponding estimator based on fallible data is

$$\hat{r}'_{yx_1} = \hat{b}'_1 \frac{s_{x_1}}{s_y} . \tag{39}$$

Since b'_1 is less than b_1, r'_{yx_1} will be less than r_{YX_1} except for sampling error. Bohrnstedt and Carter (1971) have constructed extensive tables illustrating the possible "attenuating" effects of measurement error on r_{YX_1}. Psychological researchers have been cognizant of these kinds of problems since Spearman's (1904) classic paper.

The case for the coefficient of partial correlation is directly analogous to that for the partial regression weight. Following DuBois (1957, p. 137, Eq. 80) we write

$$\hat{r}'_{yx_1 \cdot x_3} = \frac{s_{yx_1} - \hat{b}'_{x_1x_2} s_{\widetilde{yx}_2}}{\sqrt{1 - \hat{b}'_{x_1x_2} s_{x_1x_2}} \sqrt{1 - \hat{b}'_{yx_2} s_{yx_2}}} \tag{40}$$

Clearly, the observed partial correlation will be subject to the same distortions as are the partial regression weights. Cohen and Cohen (1975) have furnished several examples, which are reproduced in Table 2, of kinds of bias in partial correlations that can result from errors of measurement in the partialled variable.

Insert Table 2 about here

In all of Cohen and Cohen's examples the reliability of the pretest (the partialled variable) is .7 and the posttest and research factor reliabilities are 1.0. The partial correlations may be interpreted as the correlation of the research factor with change. In the first example the observed correlation of the research variable with change is .00 while the true or strucctural coefficient equals -.23. In other examples the observed $r_{yx_2 \cdot x_1}$ underestimates and overestimates, and in some it even has a different sign than $r_{YX_2 \cdot x_1}$. Regardless of the perspective taken the same conclusion seems to obtain: errors in variables will bias statistics based on observed scores as estimators of the underlying structural model and are likely to lead to erroneous statements about the determinants of change. Additional demonstrations of the effects of measurement error have been offered by Brewer, Campbell & Crano (1970), Campbell & Boruch (1975), Campbell & Erlebacher (1970), Evans & Anastasio (1968), Hummel-Rossi & Weinberg (1975), Kahneman (1965), Linn & Werts (1973), and Lord (1963).

REGRESSION VERSUS STRUCTURAL COEFFICIENTS

We pause briefly in our review of the effects of measurements error in studies of change to reconsider the question of formulating a structural model. An applied orientation which has a long tradition in psychology and education disagrees with the importance accorded the structural model by this reviewer (cf. Draper & Smith, 1966; Graybill, 1961; Lumsden, 1976; Marston & Borich, 1977). The position of this applied tradition is that the variables of interest are the observed scores (errors included) because decisions, predictions and evaluations are based on observed rather than true scores. When decisions are based on observed scores, there is little doubt about the validity of this position.

It should be recognized, however, that with respect to decision-making the true score has been redefined as identical to the observed score. Random fluctuations in scores are no longer regarded as error but are treated as part of the inherent variability of the predictor. Thus, if one is trying to predict the observed final status or gain or one is attempting to model economic decisions where judgments of producers and consumes are based on observed values (Johnson, 1972), then the OLS regression estimators are unbiased and consistent for the parameters of interest. This last statement is subject to one qualification: if the structural model accurately reflects the causal mechanism and a group of individuals are selected for study by some nonrandom process independent of the pretest, then the structural coefficients will provide the optimal estimates (Warren, White, & Fuller, 1974). In this context inferences from regression analysis of the observed data must be limited to randomly drawn samples. However, it has also been demonstrated that if selection into the treatment groups in an ANCOVA design is made explicitly on the basis of the pretest scores, then the observed-score ANCOVA estimates are unbiased for the structural parametes (Goldberger, 1972; Kenny, 1975; Overall & Woodward, 1976a; Rubin, 1977; Weisberg, 1979). This usually means conditional randomization where the probability of assignment to a group for each value of the pretest is explicitly determined by the experimenter. When the treatment groups pretest distributions do no overlap, we have the regression-discontinuity design which has been advocated by Campbell (1969).

In almost all social science research, particularly studies of change, the structural conception is the more appropriate. The structural model represents the causal structure or theory of the data. Research concerned with theory and hypothesis testing should be conceptualized in terms of the underlying dynamics of the behavioral processes under consideration. Hanushek and Jackson (1977) are particularly lucid on this issue: "Structural equations . . . represent the way in which we believe the observed data were generated, i.e., the underlying behavioral and stochastic processes that led to the observed data. The structural representation corresponds to the theoretical models underlying the analysis and relates to the formulation of the model where a priori information about specification or coefficient values is relevant." (Hanushek & Jackson, 1977, pp. 227-228). Furthermore, if one is interested in testing competing theories, then the structural models should be estimated since the theoretical models apply to latent or true variates and not the observed values.

In another sense the structural coefficients may be taken as more basic or fundamental than those derived from the observed-score distributions (Goldberger, 1973; Hanushek & Jackson, 1977). The parameters of the distributions of the observed scores can be expressed as functions of the structural parameters, for example, Equation 34. A change in the value of <u>one</u> structural coefficient can change the values of several or all of the observed-score coefficients. Thus, if we record changes in the observed-score estimators as different samples are drawn (e.g., males and females or 1965 and 1975) we have little way of ascertaining the component(s) of the theory on which they differ. The implications of these facts are clear: most behavioral research, particularly that called "basic" research, should be conceptualized and analyzed in terms of structural models.[5] We now take up some issues related to the estimation of structural parameters.

IDENTIFICATION REQUIREMENTS

Identification of a statistical model refers to the capability of uniquely determining the value of each hypothesized component in the model. For linear statistical models all of the information contained in the observations which is available for the estimation of parameters is contained in the variance-covariance matrix $\underline{S}_{yx}$. The number of parameters which can be identified is equal to the number of unique elements in $\underline{S}_{yx}$, which is equal to $p(p + 1)/2$ where p is the total number of variables represented in the matrix. Let us develop the case considered previously where there are fallible pretest and posttest scores related as in Equation 7-9. The covariance matrix of the observable vector (y, x_1) in terms of parameters rather than sample statistics is

$$\underline{S}_{yx} = \begin{bmatrix} b_1^2\, s_{X_1}^2 + s_v^2 + s_e^2 & b_1\, s_{X_1}^2 \\ b_1\, s_{X_1}^2 & s_{X_1}^2 + s_{u_1}^2 \end{bmatrix} . \quad (41)$$

There are three observable quantities, s_y^2, $s_{x_1}^2$, s_{yx_1}, which can be used to uniquely identify the parameters $s_{X_1}^2$, $s_{e'}^2$ and b'_1. (For present purposes we ignore the fact that information about the observed means can be used to identify b'_0 as in Equation 14.) However, the structural equations are defined in terms of five parameters. Since there are only three variances and covariances, the model cannot be identified without two independent restrictions. If the measurement error variances ($s_{u_1}^2$, $s_{u_2}^2$ and s_v^2) or the reliabilities (r_{yy} and r_{11}) are known a priori, then the coefficients can be restricted to these values leaving the three unknown parameters estimable from the observed variances and covariances.

It is easily proved that when there is only a single measure of each latent variable and no information about reliabilities available that the structural model is underidentified and a priori restrictions must be imposed in order to make the parameters estimable (Johnston, 1972; Kentall & Stuart, 1961; Mandansky, 1959; Werts, Linn & Jöreskog, 1973; Wiley, 1973). This is also illustrated by the two-predictor case described in an earlier section. The structural model given by Equations 3, 4, 6 and 22 contains eight parameters ($s_{X_1}^2$, $s_{X_2}^2$, $s_{u_1}^2$, $s_{u_2}^2$, s_v^2, s_e^2, b_1 and b_2) while there are only $3(3+1)/2 = 6$ observed variances and covariances which can be used to identify the model. Without a priori information about the error variances (or some function of them, e.g., the reliabilities) or restrictions on the model (e.g., $s_v^2 = 0$), the model is underidentified and cannot be estimated from sample data.

In general, there are three methods by which the information necessary for identification may be provided: (1) actual values or estimates of the structural parameters may be determined from previous investigations, (2) the theory may restrict some of the parameters to be zero or to equal other parameters (e.g., $s_{u_1}^2 = s_{u_2}^2$); or (3) multiple measures or indicators of the latent variables or true scores may be collected (Jöreskog, 1973; Wiley, 1973). This final alternative may be thought of as imposing a factor structure on the observations and, as a method, has excited much promising new research in psychometrics, sociology and econometrics (see Aigner & Goldberger, 1977; Goldberger & Duncan, 1973). However, in this report our exclusive concern will be with the first and second methods for identifying structural models with fallible variables.

LINEARITY CONDITIONS

Even when he model can be identified by the methods just described, research on the causes of change with fallible variables faces an additional problem before estimation can proceed. That is, rationalizing or testing the assumption that the relationship of the observed dependent variable to the observed independent variables remains linear when the underlying structural relation is linear (Cochran, 1968, 1972; Kendall & Stuart, 1961; Lindley, 1947). Here linear means linear in the X_j (straight line) rather than linear in the b_j. If the structural relation is exactly linear of the form

$$Y = b_0 + b_1X_1 + b_2X_2 + e \qquad (42)$$

and the X_j and e are independently distributed with $E(Y \; X_j) = 0$, does it follow that $Y = b'_0 + b'_1x_1 + b'_2x_2 + e$, with $E(y \mid x_j) = 0$, is also exactly linear? "The answer is, in general, no; only under certain quite stringent conditions will linearity be unimpaired" (Kendall & Stuart, 1961, p. 413).

Lindley (1947) has determined the necessary and sufficient conditions or the relation to remain linear in the narrow sense, i.e., where x and e' are not independently distributed. If we assume that the model specified by Equations 7–9 holds with the assumption that the errors are mutually and serially uncorrelated, then

$$y = \sum_j b'_j x_j \tag{43}$$

iff

$$\sum (b_j - b'_j) \frac{\partial \psi \, x_j}{\partial t_j} = \sum_j b'_j \frac{\partial \psi \, u_j}{\partial t_j} \tag{44}$$

where the ψs are Fisher's cumulant generating functions – c.g.f.s. – (logarithms of the characteristic functions) of their suffix variables (Cochran, 1968, p. 650, Eq. 8.3; Kendall & Stuart, 1961, p. 417, Ex. 29.12; Lindley 1947). Thus, when the c.f.g.s. of the X_j are multiples of the c.g.f.s. of the u_j the relation will continue to be linear. Cochran (1972) observes that "roughly speaking, this implies that u_j and X_j belong to the same class of distributions. Thus if X_j is distributed as $\chi^2 s^2$, so is u_j , . . . if X_j is normal, u_j must be normal" (p. 527).

Additional conditions are necessary if we require the x_j to be distributed independently of the residual from regression(e'), that is, to maintain linearity in the <u>fuller</u> sense. Fix (1949) proved that for the case of bivariate regression if the X, u, v and e have finite means and if the variance of either the X or u exists, then both X and u must be normally distributed in order for the observed regression to remain exactly linear.

In actual data how frequently can we expect Lindley's and Fix's conditions to hold? Cochran (1972) argues that "the forces which determine the nature of the distribution of u . . . are quite different from those that determine the nature of the distribution of the correct

[true] X. Consequently, my opinion is that in such applications even the Lindley conditions will not be satisfied, except perhaps by fluke or as an approximation . . . " (p. 528). He investigated the nature of the departure from linearity in simple bivariate cases where Lindley's conditions did not hold. His results suggest that in many situations the linear component of the observed-score regression dominates the curvilinear components, even with a relatively unreliable x_1. These findings provide some support for allowing "the ordinary theory to be used as an approximation" (Kendall, 1951, p. 24). However, Cochran was unable to obtain any general results that are exact in the bivariate case, and the nature of the departures from linearity in the multiple predictor case when Lindley's or Fix's conditions are not satisfied has not been investigated at all.

SUMMARY

This concludes our initial mathematical analysis of the problems caused by errors of measurement in investigating change with linear models. A general structural model for analyzing change has been presented. The theoretical bias and inconsistency in the observed-score regression coefficients was proved, and the harmful effects of measurement error on estimates of the squared multiple correlation, mean square error, and standardized regression weights were explicated. We described several demonstrations of how large the bias and how incorrect the resulting inferences potentially could be. The interpretation and uses of the structural coefficients were contrasted with those of the regression coefficients. We introduced the concept of identifiability and showed how it was essential to determining the estimatibility of the

structural parameters. Finally, the known conditions for the linearity of the observed-score relation when the structural relation is linear were delineated.

Statistical developments from econometrics, sociology, education, psychology, biometrics, and mathematical statistics were synthesized in this chapter. This is the first comprehensive (yet, hopefully, comprehensible) analysis of the problems caused by measurement error in linear models for analyzing growth that has been made available to educational researchers. Its purpose is to alert investigators of the harmful effects of measurement error and to furnish a detailed exposition of all the major issues. If the objective is realized, future longitudinal studies will be designed with greater care and interpreted with greater caution.

FOOTNOTES

[1] In this proposal we treat only single equation estimation techniques, so the structural coefficients for the paths between X_3 and X_1 and between X_1 and X_2 are not considered. See Hanushek and Jackson (1977) or Wiley and Harischfeger (1973) for multiequation estimation techniques for the genera path model.

[2] The superscript t will be used to designate vector and matrix transposition.

[3] It is also easy to show that b'_0 is a biased and inconsistent estimator of b_0. After b_1 has been determined, b_0 may be found using the formula (Cochran, 1968, p. 651):

$$b'_j = b_j \left(1 - s_{u_j}^2 s_{x_j}^{-\frac{1}{2}}\right) - \sum_{i \neq j}^{J} s_{x_i x_j}^{-1} s_{x_i} b_j \quad .$$

[4] It should be obvious that these statements hold quite generally and not only with respect to studies of change. Indeed, the problems of measurement error considered in this report afflict all statistical models, not just those for anlyzing change. The issue with regard to test-retest data assumes greater theoretical and methodological import because of the conceptual status of the partialled variate, i.e., <u>change</u>.

Table 1

Values of $\hat{b}'_1$ and $\hat{b}'_2$ when $b_1 = 2.0$ and $b_2 = 1.0$[a]

		$r_{X_1X_2} = +0.3$			$r_{X_1X_2} = -0.3$		
r_{22}	r_{11} =	.6	.8	1.0	.6	.8	1.0
.6	$\hat{b}'_1$	1.25	1.68	2.13	1.10	1.48	1.87
	$\hat{b}'_2$	.74	.66	.58	.44	.51	.58
.8	$\hat{b}'_1$	1.20	1.63	2.06	1.13	1.52	1.94
	$\hat{b}'_2$	.99	.89	.78	.59	.69	.78
1.0	$\hat{b}'_1$	1.15	1.57	2.00	1.15	1.57	2.00
	$\hat{b}'_2$	1.25	1.13	1.00	.75	.87	1.00

[a]Adapted from Cochran (1968, p. 657), Table 11.1.

Table 2

Effects of the Fallibility of a Partialled Variable[a]

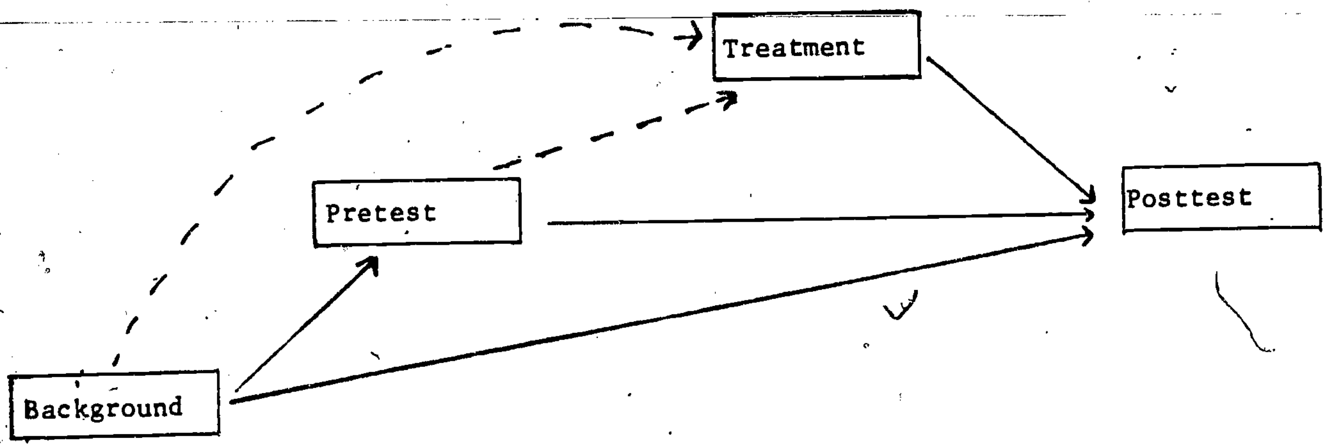
Example	r_{yx_2}	r_{yx_1}	$r_{x_1x_2}$	r_{11}	$r'_{yx_2 \cdot x_1}$	$r_{YX_2 \cdot X_1}$
1	.3	.5	.6	.7	.00	-.23
2	.5	.7	.5	.7	.24	.00
3	.5	.7	.6	.7	.14	-.26
4	.5	.3	.8	.7	.45	.57
5	.5	.3	.6	.7	.42	.37

[a]Reproduced from Cohen and Cohen, 1975, p. 371, Table 9.5.1.

Note.--For all examples, $r_{yy} = r_{22} = 1.0$.

Figure 1

General Structural Model for Studying Change[a]



[a] Adapted from Wiley and Harnischfeger (1973, p. 48)

Figure 2

Structural Model of Change including Errors of Measurement

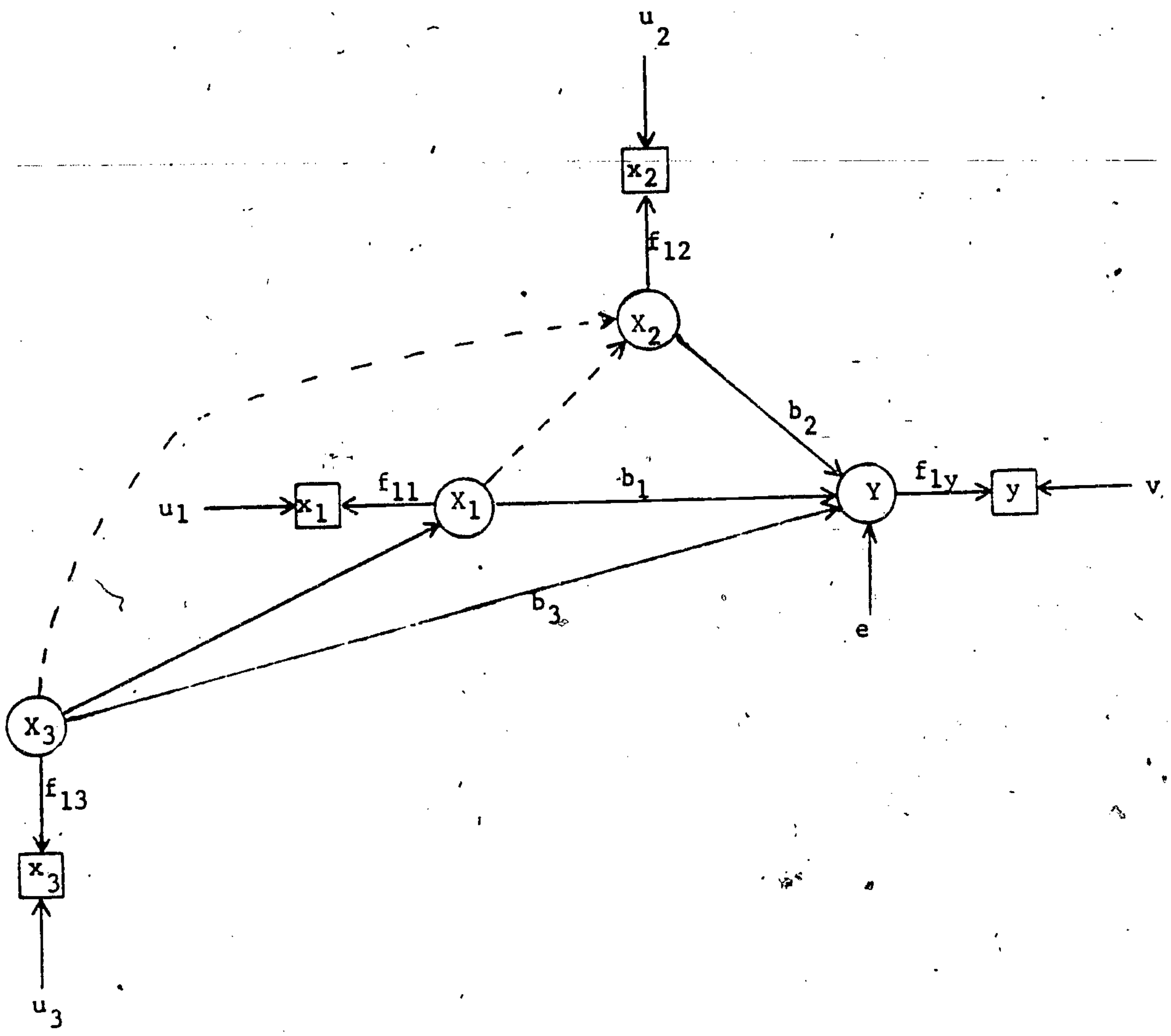
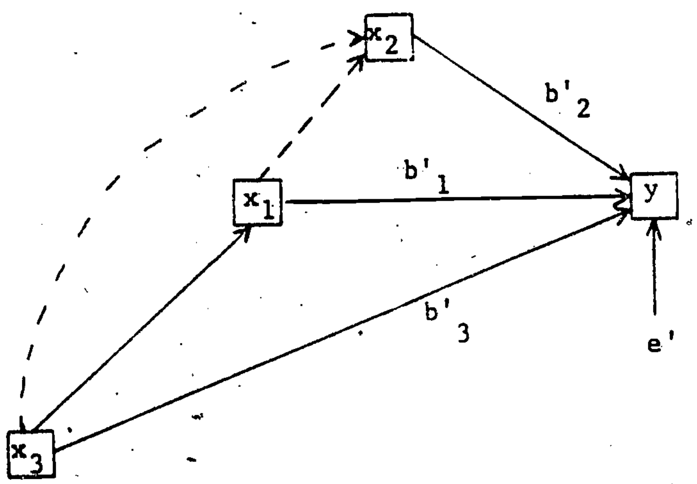


Figure 3

Typical Regression Model for Studying Change



CHAPTER III

REVIEW OF METHODS THAT CORRECT FOR ERRORS OF MEASUREMENT

INTRODUCTION

In the previous chapter we established the importance of structural models for explaining change, proved the bias of OLS regression of observed scores for estimating the parameters of the true score distributions, demonstrated the potential deleterious effects of measurement error on inferences concerning the determinants of change, and considered the requirements for identification and linearity. Now a variety of single-equation statistical methods that have been devised to estimate the structural equations can be reviewed. Explication of multiple-equation models and multiple-indicator structural equation models (e.g., Duncan & Goldberger, 1973; Aigner & Goldberger, 1976, Sorbom, 1978) lie beyond the scope of this report. (However, see Chapter IV.) In this chapter our attention will focus exclusively on techniques which utilize a priori information about the errors of measurement in the estimation process. The objective is to draw together techniques from diverse sources, to express them in a common algebra that is synchronous with the equations of the preceding chapter, and to analytically evaluate them in terms of statistical criteria, such as bias, power, and robustness. The derivations and analytic results should prove of value to educational researchers who wish to estimate the structural parameters of change. In the first section we will consider the original attenuation corrections of Spearman and then in succeeding sections four multiple regression methods that are suitable for the study of true change.

SPEARMAN'S CORRECTION FOR ATTENUATION

Although statisticians have been aware of the bias in OLS regression caused by errors of measurement for a long time (see Adcock, 1878; Kummell, 1879; Pearson, 1901), Spearman (1904) was probably the first to derive an expression for the bias and propose a method for correcting the OLS estimators for attenuation due to errors in variables. Without presenting a proof, Spearman (1904, p. 90) suggested that a zero-order correlation could be corrected for the attenuation caused by errors of measurement by dividing the observed correlation by the product of the square roots of the reliabilities:

$$\hat{r}_{YX} = \frac{\hat{r}_{yx}}{\sqrt{\hat{r}_{yy}}\sqrt{\hat{r}_{xx}}} \quad . \qquad (1)$$

In psychological research the investigator is usually interested in the association between two constructs or latent variables and not the attenuated correlation between observables. Thus, it is frequently recommended that correlations be corrected for attenuation (e.g., Block, 1963). One difficulty with Spearman's procedure, however, resides in the fact that a corrected correlation may exceed 1.0, which has given the technique a skeptical audience (A.P.A., 1966). Such an outcome may result from sampling error in the correlation or in the reliabilities. However, Bock & Petersen (1975) have developed a restricted maximum likelihood estimator of the attenuation-corrected correlation which cannot be greater than 1.0. An additional problem has been the lack of an exact formula for the standard error of a corrected correlation and procedures for hypothesis testing. Approximate formulas have been offered by Shen (1924), Cureton (1936), Kelley (1947), and Forsyth and Feldt (1969).

Utilizing the finding that the sampling distribution of a corrected correlation approximates a normal curve, Forsyth and Feldt (1969) adapted Kelley's (1947) formula to provide estimates of the standard error and to test hypotheses based on normal distribution theory. According to the results of Monte Carlo studies, their method gives reasonably good control of Type I error as indicated by the correspndence of empirical and nominal alpha levels. In addition, the procedure worked very adequately for establishing 90 and 95 per cent confidence intervals for r_{YX}. Thus, it would appear that for questions concerning the relative stability of individual differences in an attribute, Forsyth and Feldt's (1969) method can be recommended.

If one wishes, however, to test the hypothesis of no change or perfect stability in individual differences in a trait that is unreliably measured, then a different hypothesis testing strategy should be pursued. In a comparison of their normal curve procedure with a modification of McNemar's (1958) test of the hypothesis that the population correlation corrected for attenuation equals 1.0, Forsyth and Feldt (1970) found that the Forsyth-Feldt modification of McNemar's test produced empirical alpha values closer to nominal values in a series of simulated experiments (see Chapter IV). In the meanwhile Jöreskog (1971, 1974) has devised a maximum likelihood test that not only evaluates $H_o : r_{YX} = 1.0$ but evaluates the assumptinos upon which the McNemar (1958) and Lord (1957) tests are based. However, Jöreskog's covariance structure analysis requires multiple measures of time 1 and time 2 status. The conditions under which either the Forsyth-Feldt-McNemar (1970) test of Jöreskog's procedure is relatively superior have not been determined. Clearly, both of these procedures have useful roles to play in the study of change. When only single pre- and post-measurements and estimates of the reliabilities are available, the Forsyth-Feldt-McNemar test should be used. These procedures are developed more extensively in the next chapter.

When there are measurements available on treatment or background factors and interest centers on the effects of these variables on change, an index of effect which has been often recommended is the partial correlation (cf. Bereiter, 1963; Cohen & Cohen, 1975; Lord, 1963). The partial of interest is the correlation between final status and the determinant of change with the pretest partialled out. As noted above, this may be interpreted as the correlation between change and the treatment or background factor. But if it is computed from fallible observations, erroneous inferences may result. To overcome this problem, Spearman's bivariate correction for attenuation has been generalized to partial and semipartial correlations.

The formulas for a partial correlation corrected for attenuation were first presented by Stouffer (1936a, b) and have since been independently derived or discussed by Bereiter (1963), Bohrnstedt (1969), Bohrnstedt and Carter (1971), Cohen and Cohen (1975), Dunivant (1975), Hotelling (1933), Hummel-Rossi and Weinberg (1975), Kahneman (1965), Lord (1958, 1963, 1974), Meredith (1964), Mulaik (1971), O'Connor (1972), Saunders (1951), and Tucker, Damarin, and Messick (1967). To derive the formula for the estimator of the fully corrected partial correlation, one simply substitutes into the formula for the true partial correlation

$$r_{YX_2 \cdot X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{\left(1 - r_{YX_1}^2\right) \left(1 - r_{X_1X_2}^2\right)}} \tag{2}$$

the three estimators from Spearman

$$\hat{r}_{YX_1} = \frac{\hat{r}_{yx_1}}{\sqrt{(\hat{r}_{yy})(\hat{r}_{11})}}, \quad \hat{r}_{YX_2} = \frac{\hat{r}_{yx_2}}{\sqrt{(\hat{r}_{yy})(\hat{r}_{22})}}, \quad \hat{r}_{X_1X_2} = \frac{\hat{r}_{x_1x_2}}{\sqrt{(\hat{r}_{11})(\hat{r}_{22})}} \tag{3}$$

Simplifying the expression, the final result may be written

$$\hat{r}_{YX_2 \cdot X_1} = \frac{\hat{r}_{yx_2} \hat{r}_{11} - \hat{r}_{yx_1} \hat{r}_{x_1x_2}}{\sqrt{(\hat{r}_{yy} \hat{r}_{11} - \hat{r}_{yx_1}^2)(\hat{r}_{22}\hat{r}_{11} - \hat{r}_{x_1x_2}^2)}} \tag{4}$$

The formulas for incompletely corrected partial and semipartial correlations, i.e., corrected for unreliability in only one or two of the variables, are given in Dunivant (1975).

Although social researchers have been urged to use these corrected rather than the observed-score partials, they have not been given the means to place exact confidence limits on or to test hypotheses about the corrected coefficients. Lord (1974, 1975) used large sample procedures to develop the sampling theory of corrected partials. He succeeded in deriving an asymptotically efficient estimator of the corrected-for-attenuation partial correlation (which is essentially identical to the one arrived by the substitution procedure outlined in the preceding paragraph). Furthermore, Lord (1975; Stocking & Lord, 1974) provides an asymptotic test for the estimator utilizing a numerical differentiation computer program. Noting that the estimator could become infinite because of sampling fluctuations in the zero-order correlations of reliabilities, he suggests the corrected estimator "may show very large sampling fluctuations if the sample is too small, especially if either Y or X_2 is highly correlated with the true score X_1. The sampling fluctuations (of the estimator) will in certain cases be so large as to make the calculation of $r_{YX_2 \cdot X_1}$ almost useless " (Lord, 1974, p. 215).

While the observed-score partial may lead to faulty inferences about true change, it seems that the corrected-for-attenuation partial may prove little better. However, in situations where the corrected coefficient is finite and the standard error of the corrected partial is not very large as computed by the AUTEST program (Stocking & Lord, 1974), inferences should be drawn on the basis of the corrected partial. It is clear from Lord's (1974) warnings that investigators would be very unwise to compute partial correlations corrected for attenuation and fail to estimate their sampling variation. This problem of determining the sampling distributions of the estimators corrected for errors of measurement will reappear often as the other methods are discussed.

STOUFFER-LINDLEY METHOD

The first general procedure for correcting for errors of measurement in the multiple regression case was proposed by Stouffer (1936a, b) and developed more formally by Lindley (1947), who proved several theroems on the technique. Stouffer and Lindley have been followed by many others who have discovered the correction independently or who have explained the method, including Bohrnstedt (1969), Bohrnstedt and Carter (1971), Cochran (1968, 1970, 197), Cohen and Cohen (1975), Cronbach and Furby (1970), Cronbach, Rogosa, Floden and Price (1977), DuBois (1957), DeGracie (1968), DeGracie and Fuller (1972), Fuller (1975), Harnqvist (1958), Hummel-Rossi and Weinberg (1975), Johnson (1963, 1972), Kendall and Stuart (1961), Koopmans (1977), Meredith (1964), Theil (1965, 1971), Werts and Linn (1970), Wiley and Harnischfeger (1972), and Wiley and Hornik (1973). In our presentation of the method, we will follow Johnson (1963, p. 163ff) closely.

Consider the two-predictor models presented in Equations 3, 4, 6 and 22 of Chapter II where X_1 is the pretest and X_2 is some determinant of change. The following assumptions are made :

$$E(u_1) = E(u_2) = E(v) = 0 \tag{5}$$

$$E(u_1^2) = s_{u_1}^2 \quad E(u_2^2) = s_{u_2}^2 \quad E(v^2) = s_v^2 \tag{6}$$

$$u_1 \sim NID(0, s_{u_1}^2) \quad u_2 \sim NID(0, s_{u_2}^2) \quad v \sim NID(0, s_v^2) \quad e \sim NID(0, s_e^2) \tag{7}$$

$$E(u_1 e) = E(u_2 e) = E(ve) = 0 \tag{8}$$

$$E(u_1 u_2) = s_{u_1 u_2} \quad E(u_1 v) = s_{u_1 v} \quad E(u_2 v) = s_{u_2 v} \tag{9}$$

$$E(u_1 X_1) = E(u_1 X_2) = E(u_1 Y) = E(u_2 X_2) = E(u_2 Y) = E(vX_1) = E(vX_2) = E(vY) = 0 \tag{10}$$

We recall from Chapter II that

$$s_{x_1}^2 = s_{X_1}^2 + s_{u_1}^2 \tag{11}$$

$$s_{x_2}^2 = s_{X_2}^2 + s_{u_2}^2 \tag{12}$$

and note the following identities from regression theory

$$s_{x_1 x_2} = s_{X_1 X_2} + s_{u_1 u_2} \quad , \tag{13}$$

$$s_{yx_1} = b_1 s_{x_1}^2 + b_2 s_{x_1 s_2} + s_{u_1 v} \quad , \tag{14}$$

and

$$s_{yx_2} = b_1 s_{x_1 x_2} + b_2 s_{x_2}^2 + s_{u_2 v} \quad . \tag{15}$$

If we assume that X_1 and X_2 have a bivariate normal distribution, then Y, X_1 and X_2 are multinormally distributed. Maximum likelihood estimators of the population variances and covariances are given on the left sides of Equations 11, 12, 13, 14 and 15. Now if the population measurement error variances and covariances are known a priori, then Equations 11 and 12 can be substituted into 13, 14 and 15. This results in the following system of simultaneous equations for b_1 and b_2

$$\hat{b}_1(s_{x_1}{}^2 - s_{u_1}{}^2) + \hat{b}_2(s_{x_1x_2} - s_{u_1u_2}) = (s_{yx_1} - s_{vu_1}) \qquad (16)$$

$$\hat{b}_1(s_{x_1x_2} - s_{u_1u_2}) + \hat{b}_2(s_{x_2}{}^2 - s_{u_2}{}^2) = (s_{yx_2} - s_{vu_2}) \qquad (17)$$

Solution of these two normal equations produces estimators for the structural coefficients for the effects of X_1 and X_2 on true change which are identical in form to those given by Werts and Linn (1970). (To establish the correspondence, assume that the error covariances equal zero and use the identity:

$$s_{x_j}{}^2 - s_{u_1}{}^2 = r_{jj}s^2{}_{x_j}).$$

The reader will also note that the present formulation is less restrictive than Johnston's since it allows correlated errors of measurement. In actual practice, however, the extent of such correlations will be determined infrequently and typically will be assumed to be zero. The main point is that the present formulation is general enough to handle such a situation. Inserting sample y and x_j variances and covariances and known population error variances and covariances in Equations 16 and 17 and solving yields maximum likelihood estimates of b_1 and b_2. The intercept constant and residual variance from regression are found in the usual ways. Thus, no great estimation problems are encountered as long as the observations are based on sample sizes of approximately 70 or more (cf. Stroud, 1969). This will virtually insure that the sampling errors of the observed variances do not cause negative true score variance estimates.

We now present the matrix formulation for the general case. Without writing down the expressions, we assume that the assumptions in Equations 5-10 are generalized to the multivariate case. Then in matrix notation

$$\underline{\hat{S}}_{XX}\,\underline{\hat{b}} = \underline{\hat{s}}_{yx} \quad , \tag{18}$$

$$\underline{\hat{S}}_{XX} = \underline{\hat{S}}_{xx} - \underline{S}_{uu} \quad , \tag{19}$$

and

$$\underline{\hat{s}}_{YX} = \underline{\hat{s}}_{yx} - \underline{s}_{vu} \quad . \tag{20}$$

Substituting the right sides of the last two formulas for the corresponding quantities in the first equation yields

$$(\underline{S}_{xx} - \underline{S}_{uu})\underline{\hat{b}} = (\underline{\hat{s}}_{yx} - \underline{s}_{vu}) \quad . \tag{21}$$

Then the solution to the so-called normal equations is

$$\underline{\hat{b}} = (\underline{\hat{S}}_{xx} - \underline{S}_{uu})^{-1}\,(\underline{\hat{s}}_{yx} - \underline{s}_{vu}) \quad . \tag{22}$$

Assuming that the matrix inverse exists, the method will yield a unique estimate of the vector of structural coefficients. Of course, $\underline{\hat{b}}$ is a least squares estimator of $\underline{b}$ under the stated assumptions. Lindley (1947), Kendall and Stuart (1961) and Cochran (1968, 1970) present these results in a slightly different way. According to their formulation the estimate of the structural coefficients is expressed as a weighted function of observed (biased) regression weight vector, the error variances and the predictor covariances. Since

$$(\underline{S}_{xx} - \underline{S}_{uu})\underline{b} = \underline{s}_{xy} \tag{23}$$

and

$$\underline{S}_{xx}\,\underline{b}' = \underline{s}_{xy} \quad , \tag{24}$$

then

$$(\underline{S}_{xx} - \underline{S}_{uu})\,\underline{b} = \underline{S}_{xx}\,\underline{b}' \tag{25}$$

$$\underline{b} = (\underline{S}_{xx} - \underline{S}_{uu})^{-1}\,\underline{S}_{xx}\,\underline{b}' \quad . \tag{26}$$

It should be apparent that the procedures outlined by Cronbach and Furby (1970) for correcting the observed variance-covariance matrix lead to the same solution as Johnston's since $s_{x_j}{}^2 - s_{u_j}{}^2 = r_{jj}s_{x_j}{}^2$. If the covariance matrix is standardized, i.e, transformed to a correlation matrix, then $r_{jj}s_{x_j}{}^2 = (r_{jj})(1) = r_{jj}$ become the diagonal elements which is Stouffer's (1936a, b) method. If Stouffer's corrected matrix (a correlation matrix with reliabilities on the diagonal) is standardized (i.e., transformed into a correlation matrix with unities on the principal diagonal), then we have a correlation matrix corrected for attentuation by the Spearman formula given as Equation 1. Meredith (1964) described the application of multivariate statistical techniques, e.g., canonical correlation, to the attenuation-corrected correlation matrix. This method produces standardized partial regression coefficients which can be rescaled in the metric of the true scores as Cohen and Cohen (1975) describe. It should be apparent that if we calculate partial correlations from any of the corrected matrices described in this section, the coefficients will equal those calculated by the Spearman-based formulas given in the preceding section. This identity suggests that there may be problems in specifying the sampling theory for the estimators of the structural parameters computed from the observed covariance matrix corrected for errors of measurement.

There are at least three potential problems when one wishes to establish confidence limits and test hypotheses about the structural coefficients. First, sampling error in the observed variances and covariances may produce a corrected matrix which is nonGrammian and therefore not admissable as a covariance matrix (Bock & Petersen, 1975; Fuller & Hidiroglou, 1977; Williams, 1959). Second, the population error variances or reliabilities will be estimated from prior studies rather than being known. Sampling errors in the error variances can have the same damaging effect as sampling errors in the observed variances and covariances. Furthermore, this contributes another source of variability to the estimated regression coefficients which is not represented in the formulas for the standard errors of the b_j. Thus, confidence intervals and hypothesis tests will be approximte at best (Cohen & Cohen, 1975; Fuller, 1977; Warren, White, & Fuller, 1974). The third problem which may be encountered is the exacerbation of the first two problems by multicollinearity in the data, specification error in the model, etc. Although some have recommended the use of the standard formulas for calculating F- and t-ratios and standard errors (e.g., Cohen & Cohen, 1975), these potential problems should give the researcher a skeptical attitude when interpreting such statistics. It may be noted that if an expression for the standard error could be written which included information about the sampling error of the error variance or reliability, Stocking and Lord's AUTEST program could be used to provide an asymptotic test.

Before concluding this section we consider the issue of analyzing the effects of treatment or background groups on change via the Stouffer-Lindley method. As in a regular multiple regression analysis, information about group effects may be represented in a variety of ways by means of dummy-coded vectors. Analysis of covariance can be handled in this way as described earlier on; but what are consequences of including dummy-coded variables in the corrected covariance matrix? Four questions arise when we try to evaluate the effects of treatment or background group characteristics on change by means of dummy-coded variables. First, what is the consequence of violating the assumption of multivariate normality which will result from the inclusion of dummy variables? Second, can the procedure of testing for heterogeneity of regression slopes by means of product vectors be extended to this attenuation-correction method? Third, what are the effects of heterogenous error variances or reliabilities between groups on the corrected estimators? And, finally, how can error of measurement (classification) in the group factors be incorporated into the analysis? Particularly in field or quasi-experimental studies, errors of this kind may be present, e.g., racial or ethnic group membership, SES status, or culturally disadvantaged. Another way of stating this problem is that the selection rule for assigning individuals to groups may not be exactly known. Recent progress on this problem has been made by Aigner (1973), Cronbach et al. (1977), Games (1975), Mouchart (1972, 1977), and Murray (1971). To conclude this section we remark that even though the

Stouffer-Lindley method has been around for quite some time, there is still a great deal to learn about its performance in practical applications.

STROUD'S METHOD

Stroud (1968, 1972, 1974) has developed an asymptotic chi-square test of the hypothesis of equality of conditional means and variances of true scores for two groups, which is based on unrestricted maximum-likelihood estimators described by Wald (1943). The estimator of the covariance matrix for the latent variables is the same as that for the Stouffer-Lindley method. However, the covariance matrices for the two groups are each scaled in the metric of the error variance which is assumed to be the same for both groups. This implies that when the group variances on x_j differ, that the reliability of x_j for groups 1 and 2 must not be equal. This is in accord with the traditional psychometric assumption that it is the measurement error variance rather than the reliability (or, equivalently, the true score variance) which is invariant across samples of a population.

The rescaling of the covariance matrix leaves all of the scale-invariant statistics unchanged, e.g., r_{XY_1} or $\sqrt{1-r_{YX_1}^2}$; thus, the standardized results of regression analyses by the Stouffer-Lindley and Stroud methods are identical, e.g., correlations, standardized partial regression weights, and significance tests. However, the scale-dependent statistics are not unaffected by the rescaling. Therefore, the b_j or $s_{Y.X_j}^2$ will not generally agree between the two methods. This reviewer recommends that Stroud's estimates of the raw partial regression coefficients and the residual variance be rescaled to the metric of the true scores (to conform with the Stouffer-Lindley estimates) for purposes of interpretation and description.[1]

The contribution of Stroud's method is that it provides a significance test for the homogeneity of regression for two samples. Thus, it could be used to draw inferences about the differences in change between two groups. Unfortunately, the generality of the method is limited by the fact that (a) sampling error in the estimates of the variances of the measurement errors is not taken into account, (b) the method has been formulated for only the two-sample problem (although its author comments that the generalization is straightforward), and (c) separate tests of the intercept and slope parameters are not given so that tests analogous to ANCOVA's tests of differences in adjusted posttest means (or gain) are not possible. A strength of the method is that a multivariate generalization, i.e., multiple dependent variables, which uses Lord's AUTEST program, has been developed (Stroud, 1968, 1974).

FULLER'S METHODS

During the past decade Fuller and his associates have devised several procedures for correcting for errors of measurement. We will present in detail his most general formulation (Fuller, 1980) and then the related techniques more briefly.

The model and assumptions posited by Fuller (1980) for his case (i) are virtually identical to those stated for the Stouffer-Lindley method. Particularly, errors in the model (e) as well as errors in variables

(v_j, u_j) are permitted and assumed to follow a multivariate normal distribution. Furthermore, covariances between errors may be nonzero. Fuller (1980, p. 7ff) defines the estimator of the structural parameters as

$$\hat{\underline{b}} = (\hat{\underline{H}} + \frac{a}{n} \hat{\underline{S}}_{uu})^{-1} (\hat{\underline{c}} + \frac{a}{n} \hat{\underline{s}}_{uv}) \quad , \qquad (27)$$

where $a \geq 0$ is a fixed real number

$$\hat{H} = \begin{cases} \hat{\underline{S}}_{xx} - \hat{\underline{S}}_{uu} & , \text{ if } \hat{g} \geq 1 + n^{-1} \\ \hat{\underline{S}}_{xx} - (\hat{g} - n^{-1})\underline{S}_{uu} & , \text{ if } \hat{g} < 1 + n^{-1} \end{cases} \quad , \qquad (28)$$

$$\hat{\underline{c}} = \begin{cases} \hat{\underline{s}}_{yx} - \hat{\underline{s}}_{vu} & , \text{ if } \hat{g} \geq 1 + n^{-1} \\ \hat{\underline{s}}_{yx} - (\hat{g} - n^{-1})\underline{s}_{vu} & , \text{ if } \hat{g} < 1 + n^{-1} \end{cases} \quad , \qquad (29)$$

$$\hat{\underline{S}} = \begin{bmatrix} \hat{s}_y^2 & \hat{\underline{s}}_{yx} \\ \hat{s}_{xy} & \hat{\underline{S}}_{xx} \end{bmatrix} \quad , \qquad (30)$$

$\hat{g}$ is the smallest root of $\left| \hat{S} - \hat{g}I \right| = 0$, and

$$\hat{\underline{T}} = \begin{bmatrix} \hat{s}_v^2 & \hat{\underline{s}}_{vu} \\ \hat{s}_{uv} & \hat{\underline{S}}_{uu} \end{bmatrix} \quad , \qquad (31)$$

This estimator of $\underline{b}$ differs from the Stouffer-Lindley estimator in two important ways. First, the modification associated with g "guarantees that $\underline{H}$ is a positive definite matrix, that the estimator of s_e^2 is positive and the estimtor of $\underline{b}$ possesses finite variance . . . The a-modification gives an estimator that is similar to the 'k-class' estimators used in simultaneous equation estimation" (Fuller, 1980, pp. 7-8). This latter modification produces smaller MSE of the coefficients in finite samples.

These modifications represent two important advances over the Stouffer-Lindley technique. Furthermore, Fuller (1980) provides a proof of a theorem on the limiting distribution of $\underline{b}$ which demonstrates that the estimator is asymptotically normal and unbiased and specifies the covariance matrix of the estimator in the limit so that large sample confidence limits and hypothesis tests are possible. This limiting sampling distribution assumes that the population measurement error variances are known. For a more restricted model, however, which assumes uncorrelated errors and incorporates the a-modification, Fuller and Hidiroglou (1978) provide an estimator of the limiting distribution of $\hat{\underline{b}}$ which includes information about the variability of the reliability estimates if available. Under the more restrictive conditions, Fuller and Hidiroglou (1978, Theorem 1) prove that

$$n^{1/2} (\hat{\underline{b}} - \underline{b}) \xrightarrow{\mathcal{L}} N(\underline{0}, \underline{S}_{XX}^{-1} \underline{G} \underline{S}_{XX}^{-1}) \tag{32}$$

where the elements of $\underline{G}$ are functions of the true X_jX_j covariances and the ratios of error variances to total variance, e.g., $l_j = \dfrac{s_{u_j}^2}{s_{x_j}^2}$. If the $\hat{l}_j$ are arranged in a diagnoal matrix as

$$\hat{\underline{L}}_{uu} = \text{diag}\ (\hat{l}_{11}, \hat{l}_{22}, \ldots, \hat{l}_{JJ}) , \tag{33}$$

and if there is information about the sampling error of the l_j available in the form

$$n^{1/2} (\hat{l} - l) \xrightarrow{\mathcal{L}} N(\underline{0}, \underline{P}) \tag{34}$$

then

$$n^{1/2} (\hat{\underline{b}} - \underline{b}) \xrightarrow{\mathcal{L}} N(\underline{0}, \underline{S}_{XX}^{-1} \underline{G} \underline{S}_{XX}^{-1} + \underline{F} \underline{P} \underline{F}) , \tag{35}$$

where

$$\underline{F} = \text{diag}\ (b_1 s_{x_1x_1}{}^2 , b_2 s_{x_2x_2}{}^2 , \ldots, b_J s_{x_Jx_J}{}^2) . \tag{36}$$

For the first (and only) time in this review we find an expression (Equation 35) for the sampling distribution of $\hat{\underline{b}}$ which reflects the additional source of variation in $\hat{b}_j$ introduced by the imprecision in estimating r_{jj} or $s_{u_j}^{2}$. (See also corollary 3.2 in Fuller (1975, p. 127). Of all the techniques that have been considered, Fuller's methods appear to produce estimators with the most desirable properties. It is hoped that researchers involved in studying change will begin to use these estimators so that the usefulness of Fuller's methods in practical size samples can be evaluated. Fortunately, computer programs are available for performing these analyses (Hirdiroglou, Fuller, & Hickman, 1977; Wolter & Corby, 1976). A program for performing Fuller's disattenuated regression program written in the OMNITAB programming language as part of this research (Dunivant, 1978a) appears in the Appendix.

Two additional procedures which have been developed by Fuller and his associates deserve mention before leaving this section. Models of curvilinear regressions with fallible measurements have been proposed by Wolter (1974), and Wolter and Fuller (1977a, b). These methods may be useful in describing the growth curve for an attribute as a function of level of initial status. An analysis of covariance model has been developed by DeGracie (1968) and DeGracie and Fuller (1972). Their procedure contains a modification similar to the g-modification described above which guarantees the existence of the variance of the pooled-within-groups slope estimator. The estimator of the slope parameter in ordinary ANCOVA is defined for the observed variables as

$$\hat{b}'_1 = \frac{\hat{s}_{yx_1}}{\hat{s}_{x_1}^{2}} \tag{37}$$

where these are pooled-within-groups coefficients based on the fallible scores. The estimator of the structural parameter which represents the slope of the pretest on posttest regression following the Stouffer-Lindley method would be

$$\hat{b}_{1_{SL}} = \frac{\hat{s}_{YX_1}}{\hat{s}_{X_1}^2} = \frac{\hat{s}_{yx_1}}{(\hat{s}_{x_1}^2 - \hat{s}_{u_1}^2)} . \tag{38}$$

DeGracie and Fuller (1972) investigated the properties of two estimators of the slope which we present here under the assumption of $E(uv) = 0$:

$$\hat{b}'_{DF1} = \frac{\hat{s}_{yx_1}}{\hat{s}_{X_1}^2} , \tag{39}$$

where

$$\hat{s}_{X_1}^2 = \begin{cases} \hat{s}_{x_1}^2 - \hat{s}_{u_1}^2 & \text{if } \hat{s}_{x_1}^2 - \hat{s}_{u_1}^2 > (1/m)s_{u_1}^2 \\ (1/m)\hat{s}_{u_1}^2 & \text{otherwise} \end{cases}$$

and m equals N minus the number of groups minus one. The second estimator was suggested by an examination of the bias in $b_{1_{DF1}}$ and is given by

$$\hat{b}_{1_{DF2}} = \frac{\hat{s}_{yx}}{\hat{s}_{X_1}^2 + 1/m \cdot \left(2\hat{s}_{u_1}^2 + \frac{2\hat{s}_{u_1}^4}{\hat{s}_{X_1}^2} + \frac{2\hat{s}_u^4}{p\hat{s}_{X_1}^2}\right)} , \tag{40}$$

where p = a times the number of replicates per group and m (as above) is a fixed positive number. DeGracie and Fuller (1972) prove that $b_{1_{DF2}}$ has a smaller bias and smaller mean square error than $b_{1_{DF1}}$. They present an extension of classical F-ratio to test the null hypothesis of no adjusted group differences in final status or gain. Although this inferential device does not take into account the sampling error of $s_{u_1}^2$, it is probably a much more appropriate test statistic than the uncorrected F for investigations of change.

PORTER'S METHOD

The final method to be reviewed was formulated by Porter for his doctoral research in 1967 and has been cited frequently since then. The technique has been more fully elaborated in Porter and Chibucos (1974) and Olejnick and Porter (1981). A very similar method was proposed independently by Hunter and Cohen (1974).[2] Porter's method, called estimated true score analysis of covariance, has much intuitive appeal: one uses the traditional psychometric formula to estimate each individual's true pretest score and then substitutes it for the observed score as the covariate in ANCOVA.

The estimating equation for true scores for person k is given in most psychometric texts as

$$\hat{X}_k = \hat{r}_{xx} (x_k - \bar{x}) + \bar{x} \qquad , \qquad (41)$$

which is simply a linear transformation of x. Thus, for statistical procedures that are invariant to linear transformations, e.g., multiple regression, correlation, ANOVA and ANCOVA, use of either x or X produces identical results. For example, $b_{yx} = b_{YX}$ and $r_{yx} = r_{YX}$.

In ANCOVA, however, all individuals may not be sampled from the same population, so that there is not a common mean x on which to regress the observed scores. For the case of pre-existing or nonequivalent groups, unequal pretest means are usually observed and offer the possibility of regressing an individual's pretest score toward his or her <u>group</u> (h) mean:

$$\hat{X}_k = \hat{r}_{xx} (x_{hk} - \bar{x}_{h.}) + \bar{x}_{h.} \qquad (42)$$

Whenever pretest mean differences occur, Equations 41 and 42 will yield different results, and Equation 42 no longer specifies a linear transformation of x across all samples or groups represented in the study, i.e., $r_{xX} < 1.0$. Werts and Linn (1971) discuss the choice of Equations 41 and 42 as depending upon whether the group means are considered fallible. If so, Equation 41 will regress them toward the grand mean. But since doing this produces results identical to those obtaned with the observed covariable, use of Equation 41 will not allow us to estimate and test the structural parameters of interest. Thus, despite claims to the contrary, Porter's method will not produce the desired results if there are no covariate mean differences. Additionally, if the covariate means are deemed unreliable, i.e., $E(u) \neq 0$ within groups, the method of estimated true score ANCOVA fails, since it is equivalent to the observed score analysis. (It should be noted that the method in these cases will yield the appropriate estimate of b_1; however, b_2 or the estimate of adjusted mean differences will be biased and equal to the OLS estimator b'_2.)

What are the properties of the estimators obtained by Porter's method when there are pre-existing differences between the groups? First of all, the pooled-within-groups estimate of the structural slope parameter will be properly estimated as b_1. Second, since the estimated true pretest group means equal the observed group means, the estimate of group effects or differences between adjusted posttest means will correspond to the structural estimator. For the two group case Porter's method yields the following estimator of the treatment effects structural parameter:

$$\hat{t} = \bar{y}_{2.} - \bar{y}_{1.} - \hat{b}_1 (\bar{x}_{2.} - \bar{x}_{1.}) = \bar{Y}_{2.} - \bar{Y}_{1.} - \hat{b}_1 (\bar{X}_{2.} - \bar{X}_{1.}) \quad . (43)$$

The final estimator of interest in ANCOVA studies of change is the mean square error or residual variance, $s_{Y.X_j}^2 = s_e^2$. It is easily shown (e.g., Porter, 1967, p. 36) that Porter's method yields a biased estimator of the residual variance; in fact, Porter's estimator is identical to the OLS estimator $s_{e'}^2$. Since $s_{e'}^2 \geq s_e^2$ generally, the fit of the structural model will not appear as good as it really is. The hypothesis test of the between-groups factor should be conservatively biased by use of the inflated estimate of MSE. However, this bias might be offset by the increase in MSE due to estimation of r_{11}. The trade-off between the upward bias in the residual variance and the failure to include the sampling error of the reliability coefficent as a source of variation in the estimators of the structural parameters could be sufficient to make the method work reasonably well in actual research.

The results of Porter's (1967) Monte Carlo studies indicate that the empirical alpha values for the estimated true-score ANCOVA F-test of adjusted posttest means only slightly exceed the nominal alphas, e.g., an empirical alpha of .075 compared with a theoretical value of .05. This does not seem unreasonable in view of the fact that Porter's estimated true covariate scores were based on reliabilities calculated as test-retest correlations based on sample sizes of 20 to 40. POrter's Table 20 (1967, p. 100) indicates that the sampling distribution of this reliability estimate can be badly skewed for moderate values of r_{xx} and have large standard errors. It would be reasonable to infer that the variance of an ANCOVA estimator derived within the framework of the Stouffer-Lindley method, e.g., Cohen & Cohen (1975) or Cronbach & Furby

(1970), would be underestimated--sometimes substantially so--because of the failure to incorporate the sampling error of the reliability of the pretest. This fact would contribute to a liberal bias in the usual *F*-ratio based on the structural estimators, i.e., the null hypothesis of no group effects on change will be rejected too frequently.

Porter's method canot be extended directly to the multiple regression case, because there is only a single sample. In that case estimated true and observed scores would be perfectly correlated. However, something of the logic of making the reliability corrections within groups has been captured by Hunter and Cohen (1974) in their estimated true-score multiple regression analysis. Following Lord (1956), they obtain multiple regression estimates of X_1 based on x_1 and x_2 and of X_2 based on x_1 and x_2. (See Hunter & Cohen [1974, Appendix II] and Lord [1956] for the expression for weights associated with x_1 and x_2.) Although Hunter and Cohen (1974) do not develop the sampling theory for their estimators, they provide a general model that will handle curvilinear relations.

APPLICATIONS OF METHODS TO REAL DATA

In this section we review analyses of actual data made by means of one of the correction-for-attenuation methods just presented. Are inferences about factors which affect change different for the observed-score and structural models?

Several test-retest data sets have been analyzed by both the Stouffer-Lindley method and OLS regression of the observed scores. Dunivant (1977a) examined the relationships of type of nursery school program (treatment), age and sex (background characteristics) to change in sex-role identity over a nine-month interval. A composite test based on five measures of sex role identity with an estimated reliability of .75 was administered to 400 children in September and again in May. These data were submitted to a multiple regression analysis of the observed scores and to the correction-for-attenuation regression procedure described by Cohen and Cohen (1975). The results of the two analyses differed in many important respects. There was even one instance of sign reversal for one of the treatment-background factor combinations. That is, the observed score analysis indicated that one combination of factors significantly facilitated change while the structural analysis indicated that the combination inhibited growth. This example clearly demonstrates the errors of inference about change that can result from errors of measurement.

In a major reanalysis of Project Follow Through data using Cohen and Cohen's (1975) method, St. Pierre and Ladner (1977) found no effects which differed in sign between the corrected and uncorrected ANCOVAs. However, the percentage of changes in inferences about treatment effects on gain from the observed score to the corrected ANCOVA was as great as 21% when a pretest reliabilty of .6 was assumed, e.g., from null to positive, negative to null, etc. The means of the standard errors for their significance tests were all smaller than those from the uncorrected ANCOVA. They concluded that "correction of the pretest for assumed unreliability can lead to changes in the conclusions that an evaluator reaches in terms of the rank order of sponsors as well as the overall level of program effectiveness (across sponsors)" (St. Pierre and Ladner, 1977, p. 21).

Stroud (1972) used his method of comparing regressions based on fallible data to determine if the pattern of change in school achievement differed for males and females between the ninth and eleventh grades. The results of his asymptotic test for equality of regressions and conditional variances did not differ from those of the uncorrected regressions.

Several applications of Fuller's methods have been presented. Only one is directly relevant to our present concerns (see below). Rindskopf (1976) used DeGracie and Fuller's (1972) disattenuated ANCOVA method to reanalyze data from the national Head Start evaluation (Circirelli, 1969) and Glass's (1970) evaluation of ESEA Title 1 programs. When low estimates of pretest reliabilities were used, the DeGracie-Fuller method yielded different conclusions than did uncorrected ANCOVA for both evaluations. Specifically, Head Start produced significant gains in Metropolitan Readiness Test scores for black children according to the DeGracie-Fuller tests but not according to traditional analysis of covariance. The analysis of Title 1 reading scores using classical ANCOVA indicated a significant negative effect of participation. When lower-bound estimates of reliability where inserted in DeGracie and Fuller's method, however, the differences between control and treatment groups were not significant. Rindskopf's (1976) reanalyses provide another important demonstration of the potentially deleterious consequences of measurement error in drawing inferences from analyses of observed change using analysis of covariance.

The other applications of Fuller's methods are not suitable for our purposes because either they do not relate to change, or they do not report comparisons with observed score regressions. However, the examples do demonstrate that the $\hat{b}'_j$ can either underestimate or overestimate the b_j (Fuller and Hidiroglou, 1977; Warren et al., 1974). Additionally, the Warren et al. (1974) analysis of managerial role performance yielded estimates all of which had larger standard errors than those from the corresponding ordinary regression. They suggest that this result is almost always to be expected. When comparing the performances of Porter's (1967) and DeGracie and Fuller's (1972) methods with real data, Rindskopf (1976) observed that the sampling variances of the Degracie-Fuller estimators always exceeded those of Porter. These results provide some support for Warren et al.'s (1974) speculation. Contrast this with St. Pierre and Ladner's (1977) decrease in standard errors with the Stouffer-Lindley method.

Rindskopf (1976) also provided a demonstration of the use of Porter's (1967) method with Head Start and Title 1 ESEA data. He found that the corrected results from Porter's method contradicted those from classical ANCOVA in the same ways that were described above for the DeGracie-Fuller method. Porter's method appeared to be more powerful than that of DeGracie and Fuller leading Rindskopf (1976) to recommend it, especially in situations where the covariate has low reliability. This conclusion must be regarded somewhat skeptically, however, since it appears that some of Rindskopf's corrections where invalid, since they created nonGrammian covariance matrices. Although the Degracie-Fuller method insures that such impossible matrices will not be constructed, Porter's ANCOVA correction method does not. Researchers wishing to apply correction methods in order to estimate true-score effects must be careful with both the Porter and Stouffer-Lindley methods to use reliability estimates that do not generate impossible data.

Olejnick and Porter (1981) recently pointed out some important considerations in applying Porter's correction method and additional illustrations of its application. Porter and Chibucos (1974) furnish a hypothetical example in which the observed score and estimated true score ANCOVAs lead to different inferences. They conclude that the estimated true-score ANCOVA should be used to evaluate change, when the pretest is fallible and pre-existing differences between the groups obtain.

CONCLUSION

This concludes our review of the major correction-for-attenuation methods which can be used test-retest studies of change where information about the reliability of the pretest is available. We have collected and analyzed methods from statistics, education, and the social sciences. The methods of Porter (1967), Stroud (1972), and DeGracie and Fuller (1972) can be used in situations appropriate for the analysis of covariance. Of these, Porter's and DeGracie and Fuller's procedures have the more general applicability. The exactness of Stroud's method, however, strongly commends it for the two-group design. The DeGracie-Fuller method appears less powerful than Porter's but this disadvantage may be offset by reduced bias and greater safety.

For the more general multiple regression kinds of analyses (including ANCOVA), researchers may select one of the Stouffer-Lindley or Fuller methods. It seems clear that for data which conform to the usual assumptions of normality, homoscedasticity, etc., the statistical estimation and testing procedures developed by Fuller (1980), Fuller and Hidiroglou (1978), and Warren et al. (1974) will prove superior to the Stouffer-Lindley methods. Not only are Fuller's methods safer in the sense that they preclude the estimation of singular covariance matrices (of the predictor variables), they yield significance tests which are valid for finite samples. We have been unable, however, to establish analytically which technique possesses greater power. This issue is addressed in the simulation studies reported in Chapter VII.

It can be concluded from our review that most of the problems associated with estimation of the true-score regression weights have been solved by the proposed methods. The unsolved problem of greatest importance involves the unbiased estimation of the sampling variances of the disattenuated regression coefficients and the validity of associated significance tests. This chapter has helped to clarify and refine these issues, and the simulation studies reported below add further insight. Questions involving the type of reliability estimate to use and the heterogeneity of regressions constitute important problems that need to be addressed in future research.

The applications of the correction methods to real data amply illustrated the kinds of errors of inference that may have resulted from errors of measurement in previous investigations of educational change. It is hoped that the explication and evaluation of the attenuation-correction methods provided in this chapter will encourage and facilitate their use in future studies.

FOOTNOTES

[1]Note that this is in contrast to Stroud's (1972) suggestion that the appropriate b_j are those from the rescaled covariance matrix.

[2]An approximate method, which is very similar in definition to Porter's, has been proposed by Corder-Bolz (1978) and evaluated in simulation studies.

CHAPTER IV

ESTIMATION OF LINEAR FUNCTIONAL RELATIONS

INTRODUCTION

The purpose of this chapter is to analyze the problem of determining if a perfect linear relation exists among two or more variables and to review some statistical methods that have been developed to estimate and test linear functional relations. By definition, a linear functional relation (LFR) exists if the true scores on two (or more) measures are perfectly correlated. Although most of the statistical work on LFR has been done by econometricians, a problem has been investigated which is formally identical to LFR in the field of psychometrics. Psychometricians have developed several statistical tests of the hypothesis that two scales measure the same attribute except for differences in means, units of measurement, and standard errors of measurement (or reliabilities). When scales satisfy these conditions they are said to be equivalent or congeneric. As is demonstrated below, equivalent tests are related by a linear functional relation. The correlation between equivalent measures, i.e., between two variables that have a linear functional relation, when corrected for attenuation (unreliability) is 1.0. In this chapter the diverse theory and methods from econometrics, statistics, education, and psychometrics are collected, compared, and integrated. Several new results are derived for the errors-in-variables problem which should prove helpful in analyzing change occuring in measures which contain errors of measurement.

Testing hypotheses about LFR or the equivalence of measures has wide application in the analysis of change, although this has not been recognized heretofore. LFR methods could be utilized in test-retest studies which are designed to provide separate estimates of unreliability and instability in the measures (see Heise, 1969). Since LFR represents a specific model specification about the relation of the pretest and posttest scores, i.e., no stochastic error (see Fuller, 1980; Isaac, 1970), LFR methods could be used like any of the methods in the previous chapter to estimate and test hypotheses about change. This has been done previously in economics but not in educational research. Some of the LFR models could be usefully applied to the problem of inferring causal effects in cross-lag panel correlation. In the context of the general (LISREL) formulation of change in latent variables presented in the second chapter, some of the methods for assessing the equivalence of scales could be employed to evaluate the adequacy of the multi-indicator measurement models relating the observed to the true scores. Finally, LFR methods could provide valuable insight concerning the invariance of measurement metrics and validities over time and between groups in program evaluations (see Bejar, 1980).

It is hoped that this review of LFR methods and the derivation of new results will prove of value to educational researchers who are concerned with the preceding problems and statistical methods. The remainder of this section is devoted to giving a precise mathematical formulation of linear functional relations and equivalence. In the next section definitional issues concerning various types of equivalent measures are discussed. After notation and data layout conventions have been introduced, methods for determining LFR are reviewed. The methods are organized according to the type of information they require. Thus, the review is divided into procedures that require replicate measures and those that use information about the variances of the errors of measurement.

For the purpose of exposition let us assume that the following simple measurement model holds for two observed tests, x and y:

$$x = X + e_x, \text{ and} \tag{1}$$

$$y = Y + e_y \tag{2}$$

where x and y are observed scores, X and Y are unobserved true scores or latent variables, and e_x and e_y are random errors of measurement. By definition x and y are functionally related measures if the correlation between the true scores X and Y is unity, i.e., they are equivalent. Although generally x and y will be pre- and posttests, respectively, the model does not require this.

The correlation among the latent variables can be estimated in several ways. If the reliabilities of the tests are known, then the correlation between the true scores (r_{YX}) can be estimated by applying Spearman's (1904a, 1904b, 1907, 1910) correction for attenuation to the sample correlation between the observed variables ($\hat{r}_{yx}$):

$$r_{YX} = \frac{\hat{r}_{yx}}{\sqrt{(\hat{r}_{yy})\ (\hat{r}_{xx})}} \tag{3}$$

If there are multiple measures or indicators of X (say, x_1 and x_2) and Y (say y_3 and y_4), alternative estimators of r_{YX} are available. For example, Lord (1957) gives the maximum likelihood estimator as

$$r_{YX} = \frac{\hat{s}_{13} + \hat{s}_{23} + \hat{s}_{14} + \hat{s}_{24}}{4\sqrt{\hat{s}_{12}\ \hat{s}_{34}}} \tag{4}$$

(For alternative formulas based on the same covariances, see Kelley [1947] and Werts & Linn [1972].) Given a particular estimate of r_{XY}, the problem is to test the hypothesis that in the population r_{YX} = 1.0. The null and restricted alternative hypothesis may be written:

$$H_0 : r_{YX} = 1.0, \qquad (5)$$

$$H_1 : r_{YX} < 1.0. \qquad (6)$$

It is also possible to express these hypotheses as tests of restrictions placed on the linear model relating Y and X. Recognizing that Y and X may differ in their means and their scaling or units of measurements, we write

$$g_1 Y = c + g_2 X \qquad (7)$$

where g_1 and g_2 are scale coefficients and c is a constant which is a function of the differences in means between Y and X. Since g_1 and g_2 can be absorbed into a new coefficient $g = g_1/g_2$ and a new intercept defined as $a = c/g_1$, Equation 7 expresses Y as a linear transformation of X. Note that the structural model given by Equation 7 does not include a stochastic term to incorporate the effects of chance distrubances or misspecification into the model. As stated, it holds that a perfect linear relation exists between Y and X. This is referred to as a functional relation in the statistical literature (Isaac, 1970; Kendall & Stuart, 1961). The null and alternative hypotheses to be tested are:

$$H_0 : g_1Y - g_2X - c = 0, \qquad (8)$$

$$H_1 : g_1Y - g_2X - c \neq 0. \qquad (9)$$

It should be apparent that since correlations are invariant over changes in scale and origin, the null hypotheses given by Equations 5 and 8 are the same.

There is a third way of formulating the model that will prove useful in developing some of the statistical tests of equivalence. To Equation 7 we add an error term f representing random fluctuations in the fit of the model:

$$g_1Y = c + g_2X + f \tag{10}$$

If the model fits the data perfectly, the error in the equation, f, will be identically 0 for all members of the population. Thus, the hypothesis of equivalence or perfect linear relation among true scores X and Y can be evaluated by testing if the variance of f exceeds zero

$$H_0 : s_f^2 = 0, \tag{11}$$

$$H_1 : s_f^2 > 0. \tag{12}$$

The null hypothesis given in Equation 11 is the same as those in Equations 5 and 8.

A variety of different ways of testing Equations 5, and 8, and 11 under various types of assumptions have been proposed. In order to explicate these methods and their assumptions we now define equivalence and present Jöreskog's model of congeneric tests.

DEFINITION OF EQUIVALENCE AND CONGENERIC TESTS

In the development of classical test theory the concept of strictly parallel tests has played a crucial role. Strictly parallel tests by definition are tests which have equal means, equal variances, equal covariances and equal validities with respect to any criterion. It follows that parallel tests have equal standard errors of measurement and equal reliabilities. A person has the same true score on parallel tests (Gulliksen, 1950; Lord & Novick, 1968). Gulliksen (1968) argues that scores on parallel tests are completely "inter-changeable." That is, scores from one parallel test can be substituted for those for another parallel test without any loss of information whatsoever. Obviously, parallel tests satisfy the criterion for equivalence, that is, they measure the same trait except for errors of measurement. However, other tests, which do not meet the rigorous requirements for parallelism, likewise satisfy the criterion. For example, tau equivalent tests, i.e., tests which meet all of the requirements for parallelism except that they have different standard errors and consequently unequal reliabilities (Lord & Novick, 1968), measure the same underlying attribute. While tau equivalent tests measure the same construct, they do so with varying degrees of accuracy (reliability). The model for essentially tau equivalent tests goes even further by relaxing the restriction on equal means. Thus, an individual will not necessarily have equal true scores on essentially tau equivalent tests; however, all true scores on one essentially tau equivalent test will differ from those on another essentially tau equivalent test only by a constant.

The most general model for tests that measure the same attribute except for errors of measurement is Jöreskog's (1971) theory of congeneric tests. In this model almost all of the restrictions on parallel tests have been eliminated. The tests need not possess equal true means, equal true or error variances, or equal reliabilities. Individuals do not have to have equal true scores on congeneric tests. This implies that congeneric tests have different origins, or means, and different scales, or units of measurement. However, true scores on one congeneric test are a perfect linear function of true scores on another, congeneric test. Congeneric tests meet Gulliksen's (1968) criteria for "scientific equivalence," that is, they measure the same underlying attribute. In factor analytic terminology congeneric tests reflect a single general common factor. Thus, although congeneric tests are not strictly inter-changeable or substitutable and do not possess equal accuracy, they contain information about the same latent variable which underlies each of the tests. A detailed comparison of the various test models is provided in Table 1. Throughout the remainder of this chapter we shall

Insert Table 1 about here

use the terms equivalent and congeneric synonymously. For most research applications it is the congeneric type of equivalence that will be of interest. Now these ideas about the equivalence of measures is given a more precise mathematical form by developing Jöreskog's (1970, 1971, 1974) theory of congeneric tests.

Let us assume that there are two replicate measures on each of two scales x and y. By definition x_1 is congeneric with x_2, and y_3 is congeneric with y_4. While the replications may represent two alternate forms or test-retest measurements on the same test, in most cases these replications are obtained by splitting tests x and y into halves. Such split halves may or may not be parallel, but they must be congeneric for the following developments to hold. These congeneric replicates or multiple indicators are necessary in order to identify the parameters of the true and error distributions. Without the additional information provided by the multiple measurements the true score correlation cannot be estimated or tested against a hypothesized value (e.g., 1.0). The reader is referred to the text by Hanushek and Jackson (1977) for an excellent introduction to the problem of identification.

According to classical theory the observed scores can be written as a linear composite of true and error components:

$$x_1 = X_1 + e_1, \tag{13}$$

$$x_2 = X_2 + e_2, \tag{14}$$

$$y_3 = Y_3 + e_3, \tag{15}$$

$$y_4 = Y_4 + e_4 \tag{16}$$

where X_1, X_2, Y_3 and Y_4 are true scores and the e_j are random errors of measurement. Under the classical assumptions true scores are not correlated with errors, and errors are all mutually uncorrelated. Since by definition the correlation between cogeneric tests is 1.0, $r_{X_1X_2} = r_{Y_3Y_4} = 1.0$. Therefore, new random variables X and Y can be defined which are perfectly linearly related to the true scores on the individual tests X_1, X_2 and Y_3, Y_4, respectively, as follows:

$$X_1 = m_1 + b_1X, \tag{17}$$

$$X_2 = m_2 + b_2X, \tag{18}$$

$$Y_3 = m_3 + b_3X, \tag{19}$$

$$Y_4 = m_4 + b_4X, \tag{20}$$

Substitution of Equations 14 through 17 into Equations 10 through 13 yields the congeneric measurement model:

$$x_1 = m_1 + b_1X + e_1, \tag{21}$$

$$x_2 = m_2 + b_2X + e_2, \tag{22}$$

$$y_3 = m_3 + b_3Y + e_3, \tag{23}$$

$$y_4 = m_4 + b_4Y + e_4, \tag{24}$$

In matrix form this result may be written

$$
\begin{bmatrix} x_1 \\ x_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{bmatrix} + \begin{bmatrix} b_1 & 0 \\ b_2 & 0 \\ 0 & b_3 \\ 0 & b_4 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}
$$

$$
\underline{x} = \underline{m} + \underline{B}\ \underline{t} + \underline{e} \tag{25}
$$

where $\underline{x}$ is a vector of observed scores for an individual, $\underline{m}$ is a vector of means, $\underline{B}$ is a matrix of scaling coefficients, $\underline{t}$ is a vector of true scores, and $\underline{e}$ is a vector of errors of measurement.

Without loss of generality we take the true scores to be expressed in standardized forms ($\mathcal{E}(X) = \mathcal{E}(Y) = 0$, VAR $(X) =$ VAR $(Y) = 1.0$). Then the structural model relating Y to X is

$$
g_1 Y = g_2 X, \text{ or alternatively} \tag{26}
$$

$$
Y = \frac{g_1}{g_2} X = gX, \text{ and} \tag{27}
$$

the covariance matrix of the vector $\underline{t}$ is

$$
\underline{S}_{XY} = \begin{bmatrix} 1.0 & r_{XY} \\ r_{YX} & 1.0 \end{bmatrix} \tag{28}
$$

From Jöreskog's (1970) covariance structure analysis the covariance matrix of observed variables can be written as a function of the parameter matrices

$$
\begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ s_{31} & s_{32} & s_{33} & s_{34} \\ s_{41} & s_{42} & s_{43} & s_{44} \end{bmatrix} = \begin{bmatrix} b_1 & 0 \\ b_2 & 0 \\ 0 & b_3 \\ 0 & b_4 \end{bmatrix} \begin{bmatrix} 1.0 & r_{XY} \\ r_{XY} & 1.0 \end{bmatrix} \begin{bmatrix} b_1 b_2 & 0 & 0 \\ 0 & 0 & b_1 b_2 \end{bmatrix}
$$

$$+\begin{bmatrix} s_{e_1}^2 & 0 & 0 & 0 \\ 0 & s_{e_2}^2 & 0 & 0 \\ 0 & 0 & s_{e_3}^2 & 0 \\ 0 & 0 & 0 & s_{e_4}^2 \end{bmatrix}, \quad \text{or}$$

$$\underline{S}_{xy} = \underline{B}\ \underline{S}_{XY}\ \underline{B}' + \underline{S}_{ee} \qquad (29)$$

It is also possible to specify the parametric structure of the mean vector within Jöreskog's (1970) Analysis of COVariance Structure model (ACOVS). Let $\underline{D}$ be an N x 4 matrix of scores on tests x_1, x_2, y_3, and y_4 from a sample of size N, $\underline{E}$ an N x 1 matrix of ones, and $\underline{G}$ a 4 x 4 matrix specifying constraints on the mean vector. Then the population means are structured as

$$\mathcal{E}(\underline{D}) = \underline{E}\ \underline{m}'\ \underline{G}.$$

When $\underline{G}$ is the identity matrix, $\mathcal{E}(\underline{D}) = \underline{m}'$. (30)

If x and y are cogeneric, then r_{XY} will equal 1.0. That is, $g_2/g_1 = g = r_{XY} = 1.0$. Thus, testing the hypothesis that two tests x and y measure the same attribute except for differences in means, units of measurement, and errors of measurement (or that x and y are scientifically equivalent or that they have a LFR) reduces to testing the hypothesis that $r_{XY} = 1.0$. Under cogeneric assumptions the parameters b_1, b_2, b_3, b_4, m_1, m_2, m_3, m_4, s_{e_1}, s_{e_2}, s_{e_3}, s_{e_4} are free to assume any real finite values. A path diagram for the cogeneric model is presented in Figure 1. Since the true scores are standardized, the coefficient g must equal 1.0 when x and y are cogeneric.

Insert Figure 1 about here

Before concluding this section a final point should be made concerning the rank of the matrix $\underline{S}_{xy}$. If x_1, x_2, y_3, y_4 are all congeneric, then Equation 25 can be rewritten as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} t + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

$$\underline{x} = \underline{m} + \underline{bt} + \underline{e} \tag{31}$$

and Equation 29 as

$$\underline{S}_{xy} = \underline{b}\,\underline{b}' + \underline{S}_{ee}, \tag{32}$$

which is formally equivalent to a factor analytic model with one common factor. When r_{XY} is unity, the rank of $\underline{S}_{xy}$ and $\underline{S}_{XY}$ equals one. Testing the restriction that $r_{XY} = 1.0$ is equivalent to testing whether a single factor model fits the data (Gulliksen, 1968; Jöreskog, 1974). The path diagram for the one factor model is depicted in Figure 2.

Insert Figure 2 about here

DATA ORGANIZATION AND NOTATION

In order to facilitate our comparison of the several statistical procedures for determining LFR and the equivalence of measures, a common data layout and notation will be employed for all the procedures. First, we assume that there are measurements on two tests x and y which have been split into cogeneric halves: x_1 and x_2, and y_3 and y_4 as was described above. Thus, the cogeneric measurement model specified in Equations 18 through 21 holds. The scores for N persons on the four tests are organized according to the schema presented in Table 2, which also illustrates the notational conventions that have been adopted.

Insert Table 2 about here

The population covariance matrix of the vector $(x_1, x_2, x, y_3, y_4\ y)$ may be expressed in lower triangular form as

$$\underset{(x_1,\ x_2,\ x,\ y_3,\ y_4,\ y)}{\underline{S}} = \begin{bmatrix} s_{11} & & & & & \\ s_{21} & s_{22} & \text{(symmetric)} & & & \\ s_{x1} & s_{x2} & s_{xx} & & & \\ s_{31} & s_{32} & s_{3x} & s_{33} & & \\ s_{41} & s_{42} & s_{4x} & s_{43} & x_{44} & \\ s_{y1} & s_{y2} & s_{yx} & s_{y3} & s_{y4} & s_{yy} \end{bmatrix} \tag{33}$$

The estimator of $\underline{S}$ derived from a sample of size N will be denoted $\underline{S}$, the elements of which are deviation sums of squares and cross products divided by N-1 and symbolized as s. The estimator of the population correlation matrix is derived from $\underline{S}$ and may be written

$$\underset{(x_1,\ x_2,\ x,\ y_3,\ y_4,\ y)}{\underline{R}} = \begin{bmatrix} 1.0 & & & & & \\ r_{21} & 1.0 & \text{(symmetric)} & & & \\ r_{x1} & r_{x2} & 1.0 & & & \\ r_{31} & r_{32} & r_{3x} & 1.0 & & \\ r_{41} & r_{42} & r_{4x} & r_{43} & 1.0 & \\ r_{y1} & r_{y2} & r_{yx} & r_{y3} & r_{y4} & 1.0 \end{bmatrix} \tag{34}$$

We will refer to the entries in $\underline{S}\ (x_1, x_2, x, y_3, y_4, y)$ and $\underline{R}\ (x_1, x_2, x, y_3, y_4, y)$ frequently in the sections to follow. The definitions of the other vectors and matrices remain as given in the preceding sections, e.g., $\underline{m}$, $\underline{B}$, $\underline{S}_{XY}$, etc.

REVIEW OF METHODS FOR DETERMINING EQUIVALENCE

The purpose of this section is to explicate and compare seven statistical methods designed for determining if the true scores from two or more tests are perfectly linearly related. They are divided into three sets depending upon the type of information or data required. The first group contains the three best methods of those which require replicate measures of each scale: Jöreskog (1971), Kristof (1973), and Lord (1973). Gulliken (1968) and Dunivant (1979) have reviewed other less optimal procedures in this group.

In the second set are three methods which assume information is available about the covariance structure of the errors of measurement. While such information can be obtained from replicated data, it may come from any other independent sources. These methods, which were formulated primarily by statisticians concerned with estimating and testing linear functional relations, include the methods of Koopmans (1937) and Tinter (1945, 1946), Fuller (1980), and Jöreskog (1971).

The third set of methods includes only Fuller and Hidiroglou's (1978) method for testing matrix singularity when independent information about the reliabilities of the variables is available. The procedure uses the reliabilities to adjust the covariance matrix of observed scores in much the same way that the estimates of measurement error variances are utilized by the procedures in the second group. Indeed, all seven procedures are very similar in logic, if not in mathematical detail: each uses information about the covariance structure of the observed measures and errors of measurement (from replicate measurements, error variance estimates, or reliability estimates) to estimate the parameters of the linear functional relation.

To the extent possible the same outline has been followed in describing each of the methods. At the outset the statistical model and its assumptions are stated. Then the null and alternative hypotheses which are tested by the procedure are specified. Next we provide the computational formulas for the test statistic and describe how its significance is evaluated. If provided by the test developer, the estimator of r_{XY} under a true alternative hypothesis is presented. Finally, an evaluation of the test is made. For example, evidence which contradicts the validity of the test is discussed. If the efficiency of a test relative to one or more other tests is known, the superiority of the method is pointed out. Relevant Monte Carlo results, if available, are summarized. We also demonstrate that some tests differ only in computational methods, e.g., in the way the likelihood function is evaluated. They are identical statistical tests in all other respects. We begin our consideration with the methods of the first set which require replicate measurements.

Procedures Using Replications

Lord-Villegas Test

In 1973 Lord demonstrated to psychologists how a statistical procedure developed by Villegas (1964) to estimate linear functional relations could be used to test "the hypothesis that two sets of measurements differ only because of errors of measurement and because of differing origins and units of measurement" (Lord, 1973, p. 71). The assumptions of the Lord-Villegas procedure are summarized in Table 3, Insert Table 3 about here.

which is taken from Dunivant's (1979) review. The reader will note that in addition to the classical assumptions about true and error components, the model requires the errors of measurement from any pair of tests to follow a bivariate normal distribution. However, the errors of measurement for tests x_1 and x_2 may be correlated with those from y_3 and y_4. Although it is not stated by Lord, Kristof (1973) points out that the Lord-Villegas test requires that x_1 be parallel with x_2 and y_3 with y_4. Inspection of Table 3 reveals that the Lord (1957) test differs in assumptions from the Lord-Villegas method primarily in terms of which components are required to be jointly normally distributed and the correlation of errors. Also the Lord-Villegas test does not depend on sample size for its justification.

The null hypothesis tested by the Lord-Villegas procedure is precisely that the two tests x and y are congeneric or scientifically equivalent, i.e., H_0: $r_{XY} = 1.0$. The alternative is that the linear relationship between the true scores is less than perfect.

In order to perform the Lord-Villegas test we must compute three new matrices. Let us define the matrix $\underline{W}$, a within persons matrix, to be

$$\underline{W} = \begin{bmatrix} s_{x_W}^{\ 2} & s_{xy_W} \\ s_{yx_W} & s_{y_W}^{\ 2} \end{bmatrix}, \qquad (35)$$

where the elements in $\underline{W}$ are defined as

$$s_{xy_W}^{\ 2} = \sum_{i=1}^{N} \sum_{j=2}^{2} (x_{ij} - \bar{x}_{i.})^2 , \qquad (36)$$

$$s_{xy_W} = s_{yx} = \sum_{i=1}^{N} \sum_{j=1}^{2} (x_{ij} - \bar{x}_{i.}) (y_{i(j+2)} - \bar{y}_{i.}), \text{ and} \tag{37}$$

$$s_{y_W}^{\ 2} = \sum_{i=1}^{N} \sum_{j=3}^{2} (y_{ij} - \bar{y}_{i.})^2 \quad . \tag{38}$$

The reader will recall that all of the symbols are defined in Table 2 except $\bar{x}_{i.}$ and $\bar{y}_{i.}$ which equal $[(x_{i1} + x_{i2})/2]$ and $[(y_{i1} + y_{i2})/2]$, respectively.

Next the among persons sums of squares and cross products matrix $\underline{A}$ is written

$$\underline{A} = \begin{bmatrix} s_{x_A}^{\ 2} & s_{xy_A} \\ s_{yx_A} & s_{x_A}^{\ 2} \end{bmatrix}, \tag{39}$$

where

$$s_{x_A}^{\ 2} = 2 \sum_{i=1}^{N} (x_{i.} - \bar{x}_{..})^2 \quad ; \tag{40}$$

$$s_{xy_A} = s_{x_A} = 2\sum_{i=1}^{N} (\bar{x}_{i.} - \bar{x}_{..}) (\bar{y}_{i.} - \bar{y}_{..}) \quad , \tag{40}$$

$$s_{y_A}^{\ 2} = 2\sum_{i=1}^{N} (\bar{y}_{i.} - \bar{y}_{..})^2 \quad ; \tag{41}$$

Now we select the significance level at which we wish to evaluate H_0, say .05. From the tabled values of F we find the 1-05 = 95 percentile of the F distribution with N and N degrees of freedom. Finally, we evaluate the determinant[2] of matrix $\underline{C}$ of order 2:

$$|\underline{C}| = |\underline{A} - F_{95} \underline{W}| \quad . \tag{42}$$

The null hypothesis of equivalent measures is rejected at $< .05$ significance level if the determinant is positive and if both diagnoal terms are also positive, i.e., "H_0 is rejected if and only if the matrix C is positive definite"[3] (Lord, 1973, p. 71). Lord (1973) explains that the test is slightly conservative in that a true null will be rejected somewhat less often than the value of would indicate. Simulation experiments have verified this (see Dunivant, 1979). However, the method is almost as powerful as Kristof's (see below), which has the greatest power among those procedures which have been compared.

The control of Type I error, power, computational simplicity, and somewhat less restrictive assumptions of the Lord-Villegas procedure as compared with several other procedures (cf. Table 3) would seem to commend it to general use. However, Lord (1973) cautions that the procedure may be very sensitive to correlated errors within tests, i.e., when $r_{e_1e_2} \neq 0$ and $r_{e_3e_4} \neq 0$. For example, positively correlated errors will tend to increase the elements of W and consequently to decrease the probability of rejecting H_0. The extent to which this procedure is affected by violations of assumptions concerning measurement error correlations, linearity, and normality is unknown. Since educational and psychological data will often fail to satisfy such assumptions, the robustness of the Lord-Villegas test is an important question.

Kristof's Test

Kristof's (1973) method for testing if a perfect linear relation exists between the true scores X and Y represents a significant liberalization of the assumptions of the parallelism of x_1 and x_2 and of y_3 and y_4 required by most of the procedures based on replications. In fact, Table 3 shows that this method makes only three assumptions: 1) that the errors of measurement within scale x and within scale y are uncorrelated, 2) that the errors are not correlated with true scores, and 3) that the errors are multinormally distributed. Kristof's test appears to require fewer restrictive assumptions than any of the other methods.

The null hypothesis is "that two variables have perfect disattenuated correlation, hence measure the same trait except for errors of measurement. This hypothesis is equivalent to saying, within the adopted model, that true scores of two psychological tests satisfy a perfect linear relation" (Kristof, 1973, p. 101). This may be written as:

$$H_0 : a_1X + a_2Y + a_0 = 0 \quad (\text{for } a_1, a_2 \neq 0). \tag{43}$$

The reader should notice that either a_1 or a_2 will be less than zero, i.e., will be negative, under H_0. Equation 43 can be rearranged to clarify the nature of the perfect linear relation:

$$Y = \left(\frac{a_o}{a_2}\right) - \left(\frac{a_1}{a_2}\right) X \tag{44}$$

Obviously, if $\xi(Y) = \xi(X) = 0$ and $var(Y) = var(X) = 1.0$, then $-a_1/a_2 = r_{XY}$. The alternative hypothesis holds that Equation 43 is nonzero.

In order to drive Kristof's test statistic we define two new variables f_1 and f_2:

$$f_1 = a_1x_1 + a_2y_3 = a_1X_1 + a_2Y_3 + a_1e_1 + a_2e_3, \tag{45}$$

$$f_2 = a_1x_2 + a_2y_4 = a_1X_2 + a_2Y_4 + a_1e_1 + a_2e_4. \tag{46}$$

(Refer to Equations 13 through 24 for definitions of the variables.) Kristof observed that when H_0 is true, f_1 and f_2 will correlate exactly zero. Thus, H_0 can be reformulated as H_0: $r_{f_1f_2} = 0$. If $r_{f_1f_2}$ exceeds the critical value of $r_{f_1f_2}$ for a prespecified α level, then H_0 is rejected. The test is conservative according to Kristof (1973) so that "if rejection of H_0 occurs, then the true corresponding level $\bar{\alpha}$ will not exceed α, $\bar{\alpha} < \alpha$ " (p. 108). In order to test H_0, we compute the minimum possible value of $r_{f_1f_2}$ given the data subject to the restriction that $a_1, a_2 \neq 0$. Letting r_{min} be the minimum value of $r_{f_1f_2}$, compute

$$t = r_{min} \sqrt{\frac{N-2}{1 - r_{min}^2}} \qquad (47)$$

If the sample value of t exceeds the tabled value of t for N-2 df, then we reject the hypothesis of equivalence of x and y. A one-tailed test is performed because of the asymmetry of the alternative hypothesis, H_1: $r_{XY} < 1.0$.

Kristof (1973) describes several tests which are based on different assumptions concerning the parameters of the error distributions. However, we shall develop only the least restricted model. As a first step in computing r_{min} we rearrange the rows and columns in $\hat{\underline{S}}_{xy}$ to form $\hat{\underline{S}}$

$$\underline{S} = \left[\begin{array}{c|c} \underline{S}_{11} & \underline{S}_{12} \\ \hline \underline{S}'_{12} & \underline{S}_{22} \end{array}\right] = \left[\begin{array}{cc|cc} s_{11} & s_{13} & s_{12} & s_{14} \\ s_{31} & s_{33} & s_{32} & s_{34} \\ \hline s_{21} & s_{23} & s_{22} & s_{24} \\ s_{41} & s_{43} & s_{42} & s_{44} \end{array}\right]$$

Next an eigen decomposition of $\underline{S}_{12}$ is performed to yield orthogonal matrices $\underline{P}$ and $\underline{T}$ of order 2:

$$\hat{\underline{S}}_{12} = \underline{P}\,\underline{T}\,\underline{P}'$$

In the next step new matrices Q and U are found as follows:

$$\underline{Q} = \underline{S}^{.5} \quad \underline{P}'\, S_{11} \underline{P}\, \underline{T}^{-.5} = (q_{jk}) \text{ for } j = 1,2;\ k = 1,2 \text{ , and} \qquad (49)$$

$$\underline{U} = \underline{S}^{.5} \quad \underline{P}'\, S_{22} \underline{P}\, \underline{T}^{-.5} = (u_{jk}) \text{ for } j = 1,2;\ k = 1,2. \qquad (50)$$

Finally, we solve the quartic trigonometric equation

$$h_4 \cot^4 v + h_3 \cot^3 v + h_2 \cot^2 v + h_1 \cot v + h = 0, \quad (51)$$

where

$$h_4 = q_{11}u_{12} + q_{12}u_{11}, \quad (52)$$

$$h_3 = q_{11}(u_{22} - u_{11}) + u_{11}(q_{22} - q_{11}) + 4 q_{12}u_{12}, \quad (53)$$

$$h_2 = 3 [q_{12}(u_{22} - u_{11}) + u_{12}(q_{22} - q_{11})], \quad (54)$$

$$h_1 = q_{22}(u_{22} - u_{11})(q_{22} - q_{11}) - 4 q_{12}u_{12}, \quad (55)$$

$$h_0 = -q_{12}u_{22} - q_{22}u_{12}. \quad (56)$$

There will be four solutions of this equation, two of which must be real. From the largest root of the quartic we find $\hat{r}_{min}$ which is then used in the formula for t (Equation 47) to evaluate the null hypothesis.

Although Equations 46 through 56 may appear formidable, they are easily and quickly solved by standard computer programs e.g., IMSL (1979). Most computer installations will have a program for solving 4th degree polynomial equations in a single variable, say d. To solve Equation 51, let d = cot v and use the program to obtain the largest real root, say p, of

$$h_4 d^4 + h_3 d^3 + h_2 d^2 + h_1 d + h_0 = 0. \quad (57)$$

Then the corresponding root s' of Equation 51 can be obtained by using the inverse trignometric function to solve: p' = arccot p. It appears that Kristof's (1973) method represents an efficient procedure for testing if $r_{XY} = 1.0$ under very liberal assumptions. Although the test is conservative, it is valid in small sample applications. It has performed as well or better than other procedures in Monte Carlo studies. Its efficiency is especially pronounced with small sample sizes.

Thus, Kristof's (1973) procedure possesses some real advantages over those methods already reviewed. To summarize, the test does not require the parallelism of x_1 and x_2 and of y_3 and y_4 and permits some between-test error correlations. Large samples are not required to justify the validity of the test. Widely-available standard computer programs for solving polynomial equations can be used to compute the necessary test statistic quickly and inexpensively. In addition, Kristof's procedure conveys a tangible benefit to the user when x_1, x_2 and y_3, y_4 are not parallel. The degree to which Kristof's test is robust has yet to be determined, however.

Jöreskog's Test

Jöreskog's maximum likelihood method for estimating the parameters of and testing hypotheses about covariance structures provides a very flexible approach for investigating the equivalence of measures. As can be gleaned from Table 3 the method is based upon the large sample properties of maximum likelihood estimators and likelihood ratio tests subject to classical test theory assumptions and the multinormality of the observation vector (x_1, x_2, y_3, y_4). Parenthetically, we mention that the general ACOVS or COFAMM models can be defined so as to relax the classical assumption of uncorrelated errors. However, more replicate measures on x and y will be required in order to identify the model. It is obvious from Table 3 that Jöreskog's method compares favorably with the procedure of Kristof (1973) discussed in the last section. However, as will be shown below Jöreskog's technique allows greater flexibility, because it allows one to test a variety of restrictions and hypotheses.

For the purposes of this review we are interested specifically in testing two different null hypotheses within the framework of Jöreskog's

ACOVS model. The first is that $r_{XY} = 1.0$ which we now write as H_E (for equivalence). The second null is that for each variable the half tests meet the assumptions of equality of units of measurement and standard errors for parallel tests. Thus H_P symbolizes the null hypothesis that x_1 and x_2 are parallel and that y_3 and y_4 are also parallel. Of course, Jöreskog's method is completely general so that an interested investigator could test assumptions of the tau equivalence, or of the parallelism of four observed variables, i.e., Wilks (1946) test, etc. Although in this section equality constraints on the means or origins of the measures will not be considered, the reader should appreciate that Jöreskog's general COFAMM model readily permits tests about the structuring of the means (as was illustrated in a previous section). There is typically little interest in differences in means between tests, so this issue will not be pursued here. After considering the computational formulas, we shall describe how two alternative tests of H_E and H_P may be formulated and evaluated.

Jöreskog's (1970) general method for analyzing covariance structures assumes that the population covariance matrix $\underline{S}_{xy}$ has the form given in Equation (21) which is reproduced here for the reader's convenience:

$$\underline{S}_{xy} = \underline{B}\,\underline{S}_{XY}\,\underline{B}' + \underline{S}_{ee} \tag{29}$$

A covariance matrix of this structure is produced when the observed variables are structured as Equation 25 (reproduced here):

$$\underline{x} = \underline{m} + \underline{B}\,\underline{t} + \underline{e} \tag{25}$$

Three kinds of parameters may be contained in the parameter matrices $\underline{B}$, $\underline{S}_{XY}$, and $\underline{S}_{ee}$: (i) <u>fixed parameters</u> that are assigned a priori values, (ii) <u>constrained parameters</u> that are unknown, but equal to one or more other parameters, and (iii) free parameters that are unknown and unconstrained.

The problem is to find estimates of the constrained and free parameters which maximize the likelihood of the sample values given a model of the form of Equation 21. For most applications simple analytic solutions do not exist, so Jöreskog (1970) uses the numerical method of Davidon (1959) and Fletcher and Powell (1963) to maximize the likelihood function. Jöreskog argues that compared with variants of the Newton-Raphson technique, this is an efficient procedure which makes use of the derivatives of the likelihood function and the inverse of the information matrix. Actually, Jöreskog (1970) finds it more convenient to minimize a function 0, which is equivalent to maximizing the logarithm of the likelihood function L:

$$0 = \log \left| \underline{\tilde{S}}_{xy} \right| + \operatorname{tr} (\underline{\hat{S}}_{xy} \underline{\tilde{S}}_{xy}^{-1}) - \log \left| \underline{\hat{S}}_{xy} \right| - J, \qquad (58)$$

where $\underline{\tilde{S}}_{xy}$ contains the maximum likelihood estimates of $\tilde{s}_{jk}$ estimated under the model specified by Equation 21 and J is the number of observed variables. 0 is a function of the independent elements in $\underline{B}$, $\underline{S}_X$, and $\underline{S}_{ee}$. In large samples, N-1 times the minimum value of 0 is distributed as chi square and may be used to test the goodness of fit of the model. In addition approximate standard errors may be obtained for each estimated parameter from the inverse of the information matrix computed at the minimum of 0.

Hypotheses are tested in this approach by the likelihood ratio technique. The ACOVS[4] or COFAMM programs compute a chi square value for each specified model against the most general alternative, that $\underline{S}_{xy}$ is any positive definite matrix:

$$\chi^2 = -2 \ln \frac{L\ (\underline{S}_R)}{L\ (\underline{S}_F)}, \tag{59}$$

where L ($\underline{S}_R$) represents the likelihood under a given specification of fixed, free and constrained parameters (Restricted model), and L($\underline{S}_F$) is the likelihood under the assumption that $\underline{S}_{xy}$ is any positive definite matrix (Full model). According to Jöreskog (1970) it is possible to test any given model, say M_{R_1}, against a more general alternative, say M_{R_2}, by estimating and testing each one separately (against the most general alternative that $\underline{S}_{xy}$ is any p.d. matrix) and comparing their χ^2 goodness of fit values. The difference in chi square values is aysymptotically chi square distributed with degrees of freedom equal to the corresponding difference in degrees of freedom between the two models:

$$\chi_D^2 = \chi_{R_1}^2 - \chi_{R_2}^2 = N\ (O_{R_1} - O_{R_2}) \tag{60}$$

with

$$df_D = df_{R_2} - df_{R_1} \tag{61}$$

In general, the number of degrees of freedom on which any chi square test is based equals the difference in the number of parameters estimated under the full and restricted models.

With this introduction to Jöreskog's method we can now explicate the hypotheses (models) of interest in investigations seeking to determine the equivalence of measures. Following Jöreskog (1971) we suggest four models which could be tested:

$$M_1 : b_1 = b_2, b_3 = b_4, s_{e_1}^2 = s_{e_2}^2, s_{e_3}^2 = s_{e_4}^2, r_{XY} = 1.0$$

M_2 : $b_1 = b_2$, $b_3 = b_4$, $s_{e_1}^2 = s_{e_2}^2$, $s_{e_3}^2 = s_{e_4}^2$,

M_3 : $r_{\underline{XY}} = 1.0$

M_4 : $\underline{S}_{xy}$ is any p.d. matrix of rank 2 with the elements of B, $\underline{S}_{xy}$, and $\underline{S}_{ee}$ all free.

Each of these four models is tested against the most general model:

M_5 : $\underline{S}_{xy}$ is any p.d. matrix.

This series of tests is illustrated in the upper portion of Table 4 where the numbers of parameters and degrees of freedom are indicated. To test the hypothesis that the two tests x and y are equivalent (H_E: $r_{XY} = 1.0$) we could consider the goodness of fit of either Models 1 or 3. H_P, the

---

Insert Table 4 about here

---

null hypothesis that x_1, x_2 and y_3, y_4 are parallel, could be tested by Model 2. However, Jörekog (1974) maintains that "the value of χ should be interpreted very cautiously." He suggests that it is more informative to test the reasonableness of any restriction by fitting two different models, one of which contains the restriction, the other of which does not. "The differences between χ^2 values matter rather than the χ^2 values themselves" (Jöreskog, 1974).

In the lower portion of Table 4 are presented four model comparisons which yield tests of H_E and H_P. The differences in chi square values for the indicated models yeild tests of the following null hypotheses:

M_1 v. M_2: Given x_1, x_2 and y_3, y_4 are parallel, test if x and y are congeneric.

M_3 v. M_4: Given x_1, x_2 and y_3, y_4 are congeneric, test if x and y are congeneric.

M_3 v. M_1: Given x and y are congeneric, test if x_1, x_2 and y_3, y_4 are parallel.

M_4 v. M_2: Given x and y are not congeneric, test if x_1, x_2 and y_3, y_4 are parallel.

(Of course other possibilities exist, e.g. tau equivalence, and thee can be tested easily by the ACOVS or COFAMM programs.)

We observe that the test of the M_1 v. M_2 comparison is identical to Lord's (1957) test. They differ only in computing algorithms. The comparison of models 3 and 4 is comparable to Kristof's (1973) Case iii which was presented in the prior section. Although the underlying assumptions and null and alternative hypotheses are (roughly) the same, Kristof's and Jöreskog's test statistics differ considerably. In simulation experiments in which they have been compared, Kristof's method has been generally more efficient in small samples. When N exceeds 200, however, Jöreskog's procedure demonstrates greater power. The availability and ease of use of Jöreskog's COFAMM program are certainly advantages of this technique. However, serious questions about the method's sensitivity to departures from normality remain. Recently problems have been found with the Davidon-Fletcher-Powell algorithm which COFAMM uses (Lee & Jennrich, 1979). Thus, it seems premature at this point to recommend ACOVS/COFAMM as the optimal large-scale procedure.

Before closing this discussion, it is worth noting that Jöreskog's method affords the capability of testing which test model is appropriate for x_1 and x_2 and for y_3 and y_4. Identical tests of parallelism can be constructed using the methods of Wilks (1946), Votaw (1948), and Jöreskog (1970, 1971). These all produce likelihood ratio tests, but they differ in computational methods. The ease and flexibility of Jöreskog's procedure would seem to recommend it for testing assumptions about test score models, e.g., whether a set of measurements conforms to the assumptions of congeneric, essentially tau equivalent, or parallel tests.

In concluding this description of the ACOVS method we point out that it will yield a ML estimate of r_{XY} when H_E is not tenable and that it easily accommodates the analysis of several sets of congeneric tests (e.g., x, y, z) each of which has several replications (e.g, x_1, x_2, x_3, y_4, y_5, z_6, z_7, z_8). The null hypothesis of interest in this situation is H_0 : $r_{XY} = r_{XZ} = r_{YZ} = 1.0$. Finally, it is interesting to note that ACOVS or COFAMM can be used to test the hypothesis that the correlation for attenuation equals 1.0 in situations where replicate measurements on x and y are not available if the reliabilities or standard errors of x and y are known (cf. Equation 3). This will be considered in the next section.

Procedures Using Error Variances

Koopmans-Tintner Method

The credit for developing the first statistical procedure for testing the hypothesis that true scores have a perfect linear relation by using information about the covariance structure of the errors of measurement must be shared by many statisticians. I attribute the method primarily to Koopmans (1937) and Tintner (1945, 1946, 1950) because of their concern for significance testing and research application. Building primarily on the work of Rhodes (1927) and van Uven (1930), Koopmans (1937) proved the maximum likelihood properties of van Uven's (1930) weighted regression estimates of the parameters of a linear functional relation and derived approximate sampling distributions for the coefficients. This work was extended by Tintner (1945, 1946, 1950), who used a result of Hsu (1941), to derive an approximate asymptotic test of equivalence. He applied this method, which in the field of econometrics is now commonly referred to as the method of weighted regression, to problems of multicollinearity and homogeneous economic functions. As will be seen, the method shares certain identities with several multivariate techniques, most notably factor analysis and canonical correlation. Although none of its developers were concerned with the problem of equivalence of measures as defined in this chapter, the weighted regression method permits a test of the hypothesis that two or more scales differ only in means, units of measurement, and standard errors of measurement.

To explicate the procedure, we first form the covariance matrix of the total scores x and y, $\underline{S}_{xy}$, from the entries shown in Equation 33:

$$\underline{S}_{xy} = \begin{bmatrix} s_{xx} & s_{xy} \\ s_{yx} & s_{yy} \end{bmatrix} \tag{62}$$

The covariance structure of the measurement errors for the total test scores given in Equations 1 and 2 is defined as

$$\underline{S}_{ee} = \begin{bmatrix} s_{e_xe_x} & s_{e_xe_y} \\ s_{e_ye_x} & s_{e_ye_y} \end{bmatrix} \tag{63}$$

which explicitly permits correlated errors. For ease of presentation the method is illustrated for the case where there are only two (total) scales. The matrix formulation is completely general, however, and holds for any number of tests. Assuming that an estimate of $\underline{S}_{xy}$ is available and that $\underline{S}_{ee}$ is known, Equation 7 may be estimated by solving the two-matrix eigenproblem (cf. Bock, 1975):

$$(\hat{\underline{S}}_{xy} - \hat{u}_1 \underline{S}_{ee})\,\hat{\underline{g}} = 0 \tag{64}$$

The elements of the eigenvector $\hat{\underline{g}} = (\hat{g}_1, \hat{g}_2)$ corresponding to the smallest root $\hat{u}_1$ are LS/ML estimators of

$$g_1 Y = c + g_2 X \tag{7}$$

The intercept is computed by inserting mean values for X and Y in Equation 7 and solving for $\hat{c}$:

$$\hat{c} = \hat{g}_1 \overline{Y} + \hat{g}_2 \overline{X} \tag{65}$$

Under the null hypothesis (of equivalence) the quantity $(N-1)\hat{u}_1$ is approximately distributed as chi square with N-2 degrees of freedom. Anderson (1948) proved that the quantity $(\hat{u}_1 - N)/2N$ followed the unit normal distribution for large N. Approximations using the F distribution have been proposed by various authors as well. When the values of either of these test statistics exceeds the tabled values for the prespecified alpha level and appropriate degrees of freedom, the hypothesis that Y and X are equivalent is rejected.

The weighted regression method assumes that the population value of $\underline{S}_{ee}$ is used in the preceding calculations. Koopmans and Tintner both argue that using an estimate of $\underline{S}_{ee}$ will not greatly affect the validity or accuracy of the structural coefficient estimates as long as the variances of the true scores are much greater than the error dispersions, i.e., that the measures have high reliabilities. Malinvaud (1970) concurs, but cautions that this and other deductions apply only to the asymptotic distribution of the weighted regression. And, as he points out, "[u]nfortunately there seems to exist no study of the properties of this regression for finite samples." (p. 394) It is not known how efficient and robust this method is relative to those of Kristof, Lord, Jöreskog or others. The flexibility, generality, and ease of calculation make this technique potentially attractive. However, much more needs to be known about its small sample behavior in comparison to the other methods.

Fuller's Test

In a significant contribution to the weighted regression method, Fuller (1980; Warren, White & Fuller, 1974) derived a significance test for the smallest root of the determinantal equation (64) that is valid for small samples and modified the equation to improve the efficiency of the estimators of the functional relation coefficients. The methods devised by Fuller may be used with any number of variables; but, again for illustrative purposes we shall consider only two scales, x and y, as in the preceding section. Fuller assumes that the vector containing the errors of measurement are independently and identically distributed as a multivariate normal random variable with mean zero and covariance matrix $\underline{S}_{ee}$. The matrix $\hat{\underline{S}}_{sy}$ is positive definite and a consistent estimator of $\underline{S}_{xy}$. Finally, an unbiased estimator, $\hat{\underline{S}}_{ee}$, of a multiple of $\underline{S}_{ee}$ is available. Fuller (1980) presents formulas for the case where $\underline{S}_{ee}$ is known to be diagonal (the measurement errors are uncorrelated) and for the general case when $\underline{S}_{ee}$ is a positive semidefinite matrix. As described in Warren, White, and Fuller (1974) the null hypothesis given by Equation (8), that the variance of the stochastic error in the equation equals zero, can be tested for the special case of uncorrelated errors as follows:

Under the stated conditions and the null hypothesis that $s_f^2 = 0$, the distribution of the smallest root of

$$\left| \underline{\hat{S}}_{xx} - \hat{u}_1 \underline{\hat{S}}_{ee} \right| = 0 \tag{66}$$

can be approximated by Snedecor's F

$$\left(\frac{N - 1}{N - 2}\right) \hat{u}_1 \sim F \tag{67}$$

with N-2 and d degrees of freedom where

$$d = \frac{(\underline{\hat{g}}' \, \underline{\hat{S}}_{ee} \, \underline{\hat{g}})^2}{(N-1) \, \hat{g}_2^{\,2} \hat{s}_{e_x}^{\,2}} \tag{68}$$

When the obtained F exceeds the critical value of F for N-2 and d degrees of freedom, the hypothesis of equivalence or perfect linear relation is rejected.

The consistency and small sample properties of Fuller's estimators also hold for those of Koopmans and Tintner under the same assumptions. Fuller's test statistic should be better behaved than Tintner's approximations. However, the performance of these tests when the assumptions of normality and linearity are violated is unknown. The power of the weighted regression methods relative to that of Kristof's and Jöreskog's tests has not been determined either.

Jöreskog's Model

The general form of Jöreskog's method has been discussed extensively in preceding sections. Thus, it will be considered only briefly here. In Jöreskog's (1973) LISREL (LInear Structural RElations) formulation it is possible to test the hypothesis that $s_f^2 = 0$ in terms of the difference between two model chi squares. The path model is illustrated in Figure 3. This test has all of the characteristics described for the ACOVS test above and likewise depends upon assumptions of multivariate

---

Insert Figure 3 about here

---

normality and large sample size (Jöreskog & Sörbom, 1978). This LISREL test would be expected to perform very similarly to the ACOVS/COFAMM test. This is the final method which is based on information about the measurement error covariance structure to be considered. Now the only method to be considered which uses independent information about the reliabilities of x and y will be reviewed.

Procedures Using Reliabilities

Fuller-Hidiroglou Test

Fuller and Hidiroglou (1978) developed a model which uses information about the reliabilities of x and y to estimate the measurement error variances. Once the error variance estimates are obtained, the hypothesis testing procedure closely parallels Fuller's (1980) method that was presented in an earlier section. However, more stingent assumptions are required in the present case. In addition to assumptions about the normality of the error distributions, Fuller and Hidroglou (1978) assume that the true scores, X and Y, are normal independent random variables with mean zero. Although this assumption will not be tenable in some educational applications, it is not an unreasonable assumption for much of the research on educational change.

First, we define k_{yy} and k_{xx} as the ratio of error variance to total variance

$$k_{yy} = \frac{S_{e_y e_y}^2}{S_{yy}^2}, \text{ and} \tag{69}$$

$$k_{xx} = \frac{S_{e_x e_x}^2}{S_{xx}^2} \tag{70}$$

Then, the reliabilities of the observed variables may be written as

$$r_{yy} = 1.0 - k_{yy}, \text{ and} \tag{71}$$

$$r_{xx} = 1.0 - k_{xx} \tag{72}$$

Given independent estimates of the reliabilities, Equations 71 and 72 can be used to estimate k_{yy} and k_{xx}. Define $\underline{K}$ as a matrix of order two with k_{yy} and k_{xx} on the diagonal. Let $\underline{D}$ be a diagonal matrix with the standard deviations of x and y on the principal diagonal. The smallest root of the determinantal equation

$$\left| \underline{\hat{S}}_{xy} - \hat{u}_1 \underline{\hat{D}}\, \underline{\hat{K}}\, \underline{\hat{D}} \right| = 0 \tag{75}$$

may be used to test the hypothesis of equivalence. If $\hat{u}_1$ is not significantly different from one, the hypothesis of equivalence is accepted. Since the limiting distribution of $\hat{u}_1$ is the unit normal, the quantity $\alpha^{-1} N^{.5} (\hat{u}_1 - 1)$ may be compared with the tabular values of the unit normal distribution to test the hypothesis (Fuller & Hidiroglou, 1978, p. 104).

This concludes the review of methods for determining whether a linear functional relation exists or that measures are equivalent. It is easy to see how they could be used to good advantage in many studies of educational change. The optimal method will be a function of the kinds of data available and the properties of the tests and estimates. The results obtained in this chapter ought to assist researchers in choosing the best statistical LFR method for their needs.

REFERENCES FOR CHAPTER IV

Adcock, R.J. A problem in least squares. Analyst, 1878, 5, 53-54.

Aigner, D., & Goldberger, A.S. (Eds.) Latent variables in socio-economic models. Amsterdam: North-Holland, 1976.

American Psychological Association. Manual for educational and psychological tests. Washington, D.C.: Author, 1966.

Anderson, T.W. The asymptotic distribution of the roots of certain determinantal equations. Journal of The Royal Statistical Society, Series B, Supplement No. 1, 1948, 10, 190 ff.

Bejar, I.I. Biased assessment of program impact due to psychometric artifacts. Psychological Bulletin, 1980, 87, 513-524.

Blalock, H.M. Multiple indicators and the causal approach to measurement error. American Journal of Sociology, 1969, 75, 264-272.

Block, J. The equivalence of measures and the correction for attenuation. Psychological Bulletin, 1963, 60, 152-156.

Bock, R.D. Multivariate statistical methods in behavioral research. New York: McGraw-Hill, 1975.

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cronbach, L.J., & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 456-473.

Cureton, E.E. On certain estimated correlation functions and their standard errors. Journal of Experiemental Education, 1936, 4, 252-264.

Davidon, W.C. Variable metric method for minimization. A.E.C. Research and Development Report, ANL-5990 (Rev.), 1959.

Dunivant, N. Review of procedures for determining the equivalence of measures. Paper presented at meetings of the American Educational Research Association, San Francisco, April 1979.

Dunivant, N. A bibliography on statistical methods for determining the equivalence of tests and assessment procedures. Journal Supplement and Abstract Service, 1979.

Fletcher, R., & Powell, M.J.D. A rapidly converging descent method for minimization. The Computer Journal, 1963, 6, 163-168.

Forsyth, R.A., & Feldt, L.S. An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. Educational and Psychological Measurement, 1969, 29, 61-72.

Forsyth, R.A., & Feldt, L.S. Some theoretical and empirical results related to McNemar's test that the population correlation coefficient corrected for attenuation equals 1.0. American Educational Research Journal, 1970, 7, 197-207.

Fuller, W.A. Properties of some estimators for the errors-in-variables model. Annals of Statistics and Probability, 1980, in press.

Goldberger, A.S., & Duncan, O.D. Structural equation models in the social sciences. New York: Seminar Press, 1973.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Gulliksen, H. Methods for determining equivalence of measures. Psychological Bulletin, 1968, 70, 534-544.

Hanushek, E.A., & Jackson, J.E. Statistical methods for social scientists. New York: Academic Press, 1977.

Hsu, P.L. On the problem of rank and the limiting distribution of Fisher's test function. Annals of Eugenics, 1941, 11, 39ff.

Heise, D.R. Separating reliability and stability in test-retest correlation. American Sociological Review, 1969, 34, 93-101.

Hsu, T.C., & Sebatane, E.M. The effect of differences in covariate means among treatment groups upon the F-test of the analysis of covariance. Paper presented at the Annual Meeting of the American Educational Research Association, New York, N.Y., April, 1977.

IMSL, Inc. IMSL library reference manual, Vols. 1-3. Houston: Author, 1979.

Isaac, P.D. Linear regression, structural relations, and measurement error. Psychological Bulletin, 1970, 74, 213-218.

Johnson, N.L. Systems of frequency curves generated by methods of translation. Biometrika, 1949, 36, 149-176.

Johnston, J. Econometric methods. New York: McGraw-Hill, 1963.

Jöreskog, K.G. A general method for estimating a linear structural equation system. In A.S. Goldberger & O.D. Duncan (Eds.), Structural equation models in the social sciences. New York: Seminar Press, 1973.

Jöreskog, K.G. A general method for analysis of covariance structures. Biometrika, 1970, 57, 239-251.

Jöreskog, K.G., & Sörborn, D. LISREL IV - A general computer program for estimation of a linear structural equation system by maximum likelihood methods. Chicago: National Educational Resources, 1978.

Jöreskog, K.G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36, 109-133.

Jöreskog, K.G. Analyzing psychological data by structural analysis of covariance matrices. In D.H. Krantz, R.C. Atkinson, R.D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology (Vol. II). San Francisco: Freeman, 1974.

Jöreskog, K.G., & Goldberger, A.S. Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association, 1975, 70, 631-639.

Jöreskog, K.G., Gruvaeus, G.T., & van Thillo, M. ACOVS-A general computer program for analysis of covariance structures (RB 70-15). Princeton, N.J.: Educational Testing Service, 1970.

Kelley, T.L. Fundamentals of statistics. Cambridge: Harvard University Press, 1947.

Kendall, M.G., & Stuart, A. The advanced theory of statistics. Vol. 2: Inference and relationship. London: Charles Griffin, 1961.

Koopmans, T.C. Linear regression analysis in economic time series. Haarlem: De erven F. Bohn N.V., 1937.

Kristof, W. Testing a linear relation between true scores of two measures. Psychometrika, 1973, 38, 101-111.

Lee, S-Y., & Jennrich, R.I. A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. Psychometrika, 1979, 44, 99-113.

Lord, F.M. A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 1957, 22, 207-220.

Lord, F.M. Testing if two measuring procedures measure the same dimension. Psychological Bulletin, 1973, 79, 71-72.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores (with contributions by Allan Birnbaum). Reading, Mass: Addison-Wesley, 1968.

Madansky, A. The fitting of straight lines when both variables are subject to error. Journal of the American Statistical Association, 1959, 54, 173-205.

Malinvand, E. Statistical methods of econometrics (2nd rev. ed.). New York: American Elsevier, 1970. (Tr. by Mrs. A. Silvey.)

McNemar, Q. Attenuation and interaction. Psychometrika, 1958, 23, 259-265.

Mendenhall, W., & Scheaffer, R.L. Mathematical statistics with applications. North Scituate, Mass.: Duxbury Press, 1973.

Moran, P.A.P. Estimating structural and functional relationships. Journal of Multivariate Analysis, 1971, 1, 232-255.

Nesselroade, J.R., & Reese, H.W. Life-span developmental psychology: Methodological issues. New York: Academic Press, 1973.

Nunnally, J.C. Psychometric theory (2nd ed.). New York: McGraw-Hill, 1978.

Rhodes, E.C. On lines and planes of closest fit. Philosophical Magazine, Series 7, 1927, 3, 357 ff.

Spearman, C. The proof and measurement of association between two things. American Journal of Psychology, 1904, 15, 72-101. (a)

Spearman, C. General intelligence, objectively determined and measured. American Journal of Psychology, 1904, 15, 201-293. (b)

Spearman, C. Demonstration and formulae for true measurement of correlation. American Journal of Psychology, 1907, 18, 161-169.

Spearman, C. Coefficient of correlation calculated from faulty data. British Journal of Psychology, 1910, 3, 271-295.

Tintner, G. A note on rank, multicollinearity and multiple regressions. Annals of Mathematical Statistics, 1945, 16, 304 ff.

Tintner, G. Multiple regression for systems of equations. Econometrica, 1946, 14, 5 ff.

Tinter, G. A test for linear relations between weighted regression coefficients. Journal of the Royal Statistical Society, Series B, 1950, 12, 273 ff.

Villegas, C. Confidence region for a linear relation. Annals of Mathematical Statistics, 1964, 35, 780-788.

Votaw, D.F., Jr. Testing compound symmetry in a normal multivariate distribution. Annals of Mathematical Statistics. 1948, 19, 447-473.

van Uven, M.J. Adjustment of N points (in n-dimensional space) to the best linear (n-1)-dimensional space: I and II. Koninklijke Akademie van Wetenschapen te Amsterdam, Proceedings of the Section of Science, 1930, 23, 143 ff., 307 ff.

Warren, R.D., White, J.K., & Fuller, W.A. An errors-in-variables analysis of managerial role performance. Journal of the American Statistical Association, 1974, 69, 886-893.

Werts, C.E., & Linn, R.L. Corrections for attenuation. Educational Psychological Measurement, 1972, 32, 117-127.

Werts, C.E., Linn, R.L., & Jöreskog, K.G. Another perspective on "Linear regression, structural relations, and measurement error." Educational and Psychological Measurement, 1973, 33, 327-332.

Wilks, S.S. Sample criteria for testing equality of means, equality of variance and equality of covariances in a normal multivariate distribution. Annals of Mathematical Statistics, 1946, 17, 257-281.

Wonnacott, T.H., & Wonnacott, R.J. Introductory statistics (2nd ed.). New York: Wiley, 1972.

FOOTNOTES

[1]In this chapter a prime indicates vector and matrix transposition.

[2]The determinant of a 2x2 matrix is equal to the product of the diagnoal elements minus the product of the off-diagnoal elements.

[3]In this quotation and all others cited, symbols have been changed to conform to the notational conventions used in this paper.

[4]The most recent version of Jöreskog's program for the analysis of covariance structures COnfirmatory Factor Analysis with Model Modification (COFAMM) is marketed through International Educational Resources, Inc., Box A 3650, Chicago, Illinois 60690.

Table 1

Comparison of Test Score Models[a]

Test Score Model	Propensity Distributions: First Two Moments	Propensity Distributions: Higher Moments	Experimental Independence	Linear Experimental Independence	True Scores	Error Variances	Observed Means	Intercor-relations	Validities
Strictly Equivalent	Equal	Equal	Yes	Yes	Equal	Equal	Equal	Equal	Equal
Parallel	Equal	Unequal	No	Yes	Equal	Equal	Equal	Equal	Equal
T- equivalent	Unequal	Unequal	No	Yes	Equal	Unequal	Equal	Unequal	Unequal
Essentially T- equivalent	Unequal	Unequal	No	Yes	Unequal[b]	Unequal	Unequal	Unequal	Unequal
Congeneric	Unequal	Unequal	No	Yes	Unequal[c]	Unequal	Unequal	Unequal	Unequal

a See Lord and Novick (1968, Ch. 2) for more information

b True scores may differ only by an additive constant

c True scores may differ only by an additive constant and a scaling factor

Table 2

Score Schema[a]

	Test					
	x Replication			y Replication		
Individual	(1)	(2)	Sum	(3)	(4)	Sum
1	x_{11}	x_{12}	$x_{1\cdot}$	y_{13}	y_{14}	$y_{1\cdot}$
2	x_{21}	x_{22}	$x_{2\cdot}$	y_{23}	y_{24}	$y_{2\cdot}$
.						
.						
.						
i	x_{i1}	x_{i2}	$x_{i\cdot}$	y_{i3}	y_{i4}	$y_{i\cdot}$
.						
.						
.						
N	x_{N1}	x_{N2}	$x_{N\cdot}$	y_{N3}	y_{N4}	$y_{N\cdot}$
Mean	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\bar{x}_{\cdot\cdot}$	$\bar{y}_{\cdot 3}$	$\bar{y}_{\cdot 4}$	$\bar{y}_{\cdot\cdot}$

[a]Adapted from McNemar (1958, p. 259).

Table 3

Comparison of Assumptions and Hypotheses of Eight Methods for Determining Equivalence

Assumption	Wilks-Votaw[a]	McNemar[a]	Forsyth-Feldt-McNemar[a]	Forsyth-Feldt[a]	Lord[a]	Lord-Villegas[a]	Kristof[a]	Jöreskog[a]
Large sample test	Y	N	N	Y	Y	N	N	Y
x_1 and x_2 same origins	N/H	Y	Y	Y	Y	Y	N	N/H
x_1 and x_2 same units of measurement	H	Y	Y	Y	Y	Y	N	N/H
x_1 and x_2 same std error of measurement	H	Y	Y	Y	Y	Y	N	N/H
y_3 and y_4 same origins	H	Y	Y	Y	Y	Y	N	N/H
y_3 and y_4 same units of measurement	H	Y	Y	Y	Y	Y	N	N/H
y_3 and y_4 same std error of measurement	H	Y	Y	Y	Y	Y	N	N/H
x_1, x_2, y_4 same origins	H	Y	Y	N	N	N	N	N/H
x_1, x_2, y_3 and y_4 same units of measurement	H	Y	Y	N	N	N	N	N/H
x_1, x_2, y_3 and y_4 same std error of measurement	H	Y	N	N	N	N	N	N/H
x and y same reliabilities	H	Y	N	N	N	N	N	N/H
$r_{XY} = 1.0$	H	H	H	H	H	H	H	H
$\mathcal{E}(e_1e_2) = 0$, $\mathcal{E}(e_3e_4) = 0$	Y	Y	Y	Y	Y	Y	Y	Y/H
$\mathcal{E}(e_jX) = \mathcal{E}(e_jY) = 0$	Y	Y	Y	Y	Y	Y	Y	Y/H
$\mathcal{E}(e_1e_3) = \mathcal{E}(e_1e_4) = \mathcal{E}(e_2e_3) = \mathcal{E}(e_2e_4) = 0$	Y	Y	Y	Y	Y	N	N	Y/H
$\mathcal{E}(e_1) = \mathcal{E}(e_2) = \mathcal{E}(e_3) = \mathcal{E}(e_4) = 0$	Y	Y	Y	Y	Y	Y	N	Y
x_1, x_2, y_3 and y_4 multivariate normal	Y	N	N	N	Y	N	N	Y
e_1, e_2, e_3 and e_4 normally distributed	N	Y	Y	N	N	Y	Y	N
e_1e_3, e_1e_4, e_2e_3 and e_2e_4 bivariate normal	N	N	N	N	N	Y	Y	N
e_1e_3 and e_1e_4 same joint distributions	N	N	N	N	N	Y	N	N
e_2e_3 and e_2e_4 same joint distributions	N	N	N	N	N	Y	N	N
Easily generalized to three or more tests (x, y, z,...)	Y	Y	N	N	N	N	N	Y
x and y bivariate normal distribution	N	N	N	Y	Y	N	N	N

[a] The letters make the following designations: Y - Yes, the assumption is required; N - No, the assumption is not required; H - Hypothesis, the assumption is tested as the null hypothesis.

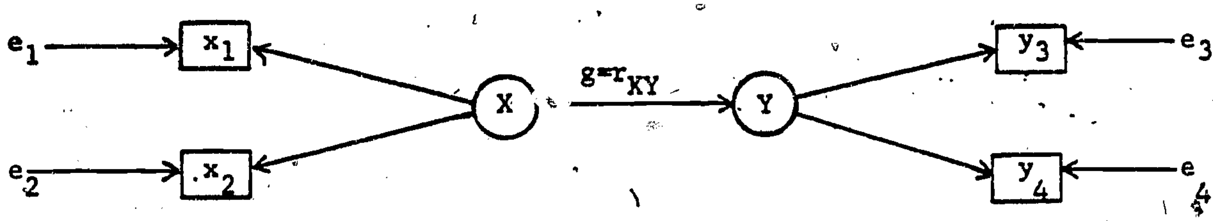
Table 5

Tests of Equivalence Using ACOVS

Restricted Model	Number of Parameters	Full Model	Number of Parameters	df for χ^2_D
M_1	4	M_5	10	6
M_2	5	M_5	10	5
M_3	8	M_5	10	2
M_4	9	M_5	10	1
M_5	10	--Not tested--		
M_1	4	M_2	5	1
M_3	8	M_4	9	1
M_3	2	M_1	6	4
M_4	1	M_2	5	4

Figure 1

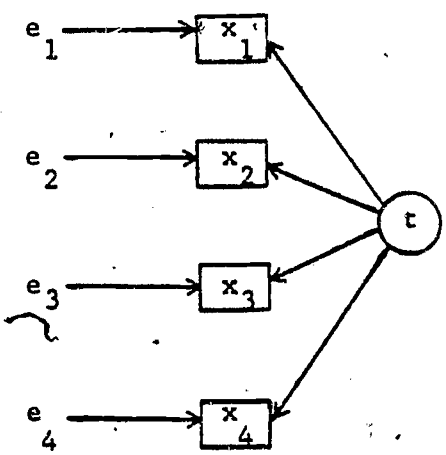
Path Model with Two Sets of Congeneric Tests[a]



[a]Adapted from Jöreskog (1974), Figure 3.

Figure 2

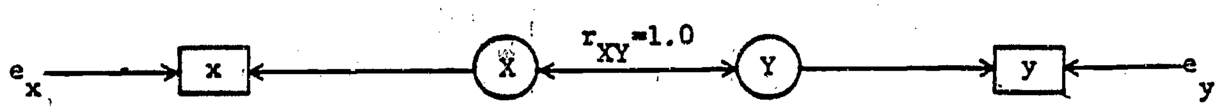
Path Model with Four Congeneric Tests[a]



[a]Adapted from Jöreskog (1974), Figure 1.

Figure 3

Path Model for Correlation Corrected for Attenuation



CHAPTER V

SOME ANALYTIC RESULTS FOR PARAMETERS AFFECTING BIAS IN GOODNESS OF FIT AND SAMPLING DISTRIBUTION STATISTICS

INTRODUCTION

The developments in the preceeding chapters suggest that the parameters of the observed-score distributions are functions of the parameters of the latent-variable distributions. This is indeed the case. We can write expressions for $\underline{b}'$, $R^{2'}$, and $s_{e'}^2$ in terms of $\underline{b}$ and the population variances and covariances of the latent (true and error) variables. In addition, for a fixed preselected sample size (N) the expected values of $\underline{S}_{\underline{bb}}$ and $\underline{S}_{\underline{b'b'}}$, the covariance matrices of $\underline{b}$ and $\underline{b}'$, can be derived in terms of the structural parameters. We present these results in this chapter and compare $R^{2'}$ with R^2, $s_{e'}^2$ with s_e^2, and $\underline{S}_{\underline{b'b'}}$ with $\underline{S}_{\underline{bb}}$. The comparisons enable us to draw conclusions concerning the parameters affecting bias in the observed-score statistics. We describe the kinds of data and conditions which are likely to lead to incorrect inferences concerning the determinants of true change from observed-score regressions.

These results mean that if an investigator had hypotheses or knowledge about the structural parameters, then he or she could determine the corresponding parameter values for the observed-score population. By comparing these two sets of parameters the researcher could ascertain the degree to which inferences about true change based on analyses of (even very large) samples of observed scores could be expected to be incorrect. However, most investigators are not able to state a priori the population parameter values of the true and error distribution because of a lack of previous research or because the mathematical formalization of the verbal theory can not be accomplished precisely. Even though the exact values of the latent variable parameters are not available in most circumstances, a range of likely or theoretically possible values usually can be prespecified. For these cases, sets of possible latent structure parameter values could be used to generate sets of possible observed-score outcomes. These could be evaluated and the potential for errors of inferences due to errors of measurement assessed. In the next chapter we use the results of this chapter to devise an algorithm which takes as input the parameter values of the structural relations among the latent variates as specified by the researcher and outputs the expected values of the observed-score regression parameters for a given sample size.

EXPRESSIONS FOR THE TRUE-SCORE PARAMETERS

Before the expressions for the observed-score parameters can be written in terms of the latent-variable parameters, it is necessary to derive the covariance structure of the latent variables. First, recall the single-equation structural model specifying the true posttest (Y) as a function of the structural regression coefficients (b_0, $\underline{b}$), the vector of true causal variables ($\underline{X}$), and the stochastic error component (e) given in Chapter II as

$$Y = b_0 + \underline{b}^t\underline{X} + e \tag{1}$$

(where the superscript t represents vector or matrix transposition). The equations of the simplified measurement model are also reproduced for the reader's convenience:

$$\underline{x} = \underline{X} + \underline{u} \tag{2}$$

$$y = Y + v \tag{3}$$

Now the covariance structure of the latent variables can be given. For the true scores we have the covariance matrix of the X_j

$$\underline{S}_{XX} = \begin{bmatrix} s_{X_1}{}^2 & & & \\ s_{X_2X_1} & s_{X_2}{}^2 & \text{(symmetric)} & \\ \vdots & & \ddots & \\ s_{X_kX_1} & \cdots & & s_{X_k}{}^2 \end{bmatrix} , \tag{4}$$

and the covariances of the true X_j with the true Y

$$\underline{s}_{YX} = \begin{bmatrix} s_{YX_1} \\ s_{YX_2} \\ \vdots \\ s_{YX_k} \end{bmatrix} = \underline{S}_{XX}\, \underline{b} . \tag{5}$$

The vector of structural regression coefficients can also be written as a linear function of the true variance and covarianes as demonstrated in Chapter II:

$$\underline{b} = \underline{S}_{XX}{}^{-1}\, \underline{s}_{yx} \tag{6}$$

and then the intercept coefficient is

$$b_0 = \bar{Y} - \underline{b}^t\, \bar{\underline{X}} \tag{7}$$

where the bars designate means or expected values. Since Y is a weighted linear function of $\underline{X}$ and e, the variance of Y can be expressed as a weighted linear combination of the variances and covariances of the X_j and e, where we make explicit the usual assumption that $E(\underline{X}e) = \underline{0}$:

$$s_y{}^2 = \underline{b}^t \underline{s}_{YX} + s_e{}^2 = \underline{b}^t \underline{S}_{XX}\, \underline{b} + s_e{}^2 . \tag{8}$$

Equations 4 through 8 may be summarized in the form of the partitioned covariance matrix for Y, $\underline{X}$ as follows:

$$\underline{S}_{YX} = \left[\begin{array}{c|c} s_Y^2 & \underline{s}_{YX}^t \\ \hline \underline{s}_{YX} & \underline{S}_{XX} \end{array}\right] = \left[\begin{array}{c|c} \underline{b}\,\underline{S}_{XX}\,\underline{b}^t + s_e^2 & \underline{b}^t{}_{XX}\,\underline{S} \\ \hline \underline{S}_{XX}\,\underline{b} & \underline{S}_{XX} \end{array}\right] . \quad (9)$$

s_e^2 for True Change

The first index of the magnitude of the systematic relation between Y and $\underline{X}$, the square of the standard error of estimate (s_e^2), can be expressed in terms of the structural regression coefficients and the true variances and covariances. If we let $\hat{Y}$ represent the systematic or predictable part of Y, then

$$\hat{Y} = b_o + \underline{b}^t\underline{X} \quad (10)$$

Note that the structural equation model adopted in this paper (Equation 1) specified that $\hat{Y} \neq Y$ in the population. Thus, e, defined as

$$e = Y - (b_o + \underline{b}^t\underline{X}) = Y - \hat{Y} \quad , \quad (11)$$

can be taken as a stochastic component representing the fact that the response process or response generating mechanism is probabilistic in nature. Alternatively e can be conceived as a lack of complete model specification as follows. We take e to be a linear combination of additional predictors of Y

$$e = X_{k+1} + \dots + X_p \quad , \quad (12)$$

where regression weights are ignored and impose the restriction that $E(X_iX_j) = 0$ for $i = 1 \dots k$, $j = k + 1 \dots p$. Then e will function as a random variable in the structural model. The variance of e (the square of the standard error of estimate) is

$$s_e^2 = E(ee) = E[\{Y - (b_o + \underline{b}^t\underline{X})\}^t \{Y - (b_o + \underline{b}^t\underline{X})\}] \quad . \qquad (13)$$

Evaluating the right-hand member leads to an expression in which s_e^2 is given as a variance of a difference in terms of its components:

$$s_e^2 = s_Y^2 + \underline{b}^t\underline{S}_{XX}\,\underline{b} - 2\underline{b}^t\underline{s}_{YX} \quad . \qquad (14)$$

R^2 for True Change

The second index of the degree of systematic relation between Y and $\underline{X}$ is the coefficient of multiple determination or squared multiple correlation. It is defined as the ratio of explained variance to total variance:

$$R^2 = \frac{\underline{b}^t\underline{S}_{XX}\,\underline{b}}{s_Y^2} \quad . \qquad (15)$$

$\underline{S}_{bb}$ for True Change

For a fixed sample size the sampling variability of the regression coefficients can be derived for the general case:

$$\underline{S}_{\underline{b}\,\underline{b}} = s_e^2 \; \frac{1}{N} \; \underline{S}_{XX}^{-1} \quad . \qquad (16)$$

The right member in this equation contains information about the variance structure of Y and $\underline{X}$. Having derived a set of equations which involve parameters that apply to true-score regression, we can now focus upon the errors of measurement.

The variance structure of the predictor measurement errors will be denoted as

$$\underline{S}_{uu} = \begin{bmatrix} s_{u_1}^2 & & & \\ s_{u_2u_1} & s_{u_2}^2 & (\text{Sym}) & \\ \vdots & & & \\ s_{u_ku_1} & & & s_{u_k}^2 \end{bmatrix}, \quad (17)$$

The variance of the dependent variable error of measurement is s_v^2 and the covariance vector of v, $\underline{u}$ is

$$\underline{s}_{vu}^t = (s_{vu_1}\; s_{vu_2}\; \cdots\; s_{vu_k}) \; . \quad (18)$$

The results are amalgamated into a partitioned matrix, $\underline{S}_{vu}$:

$$\underline{S}_{vu} = \begin{bmatrix} s_{v_1}^2 & & & \\ s_{u_1v} & s_{u_1}^2 & (\text{symmetric}) & \\ \vdots & \vdots & \ddots & \\ s_{u_kv} & s_{u_ku_1} & \cdots & s_{u_k}^2 \end{bmatrix}, \quad (17)$$

EXPRESSIONS FOR THE OBSERVED-SCORE PARAMETERS

In deriving the covariance structure of the observed variables it is necessary to impose certain restrictions usually associated with classical test theory (Lord & Novick, 1968), viz., that the errors of measurement are uncorrelated with the true scores, that the expected values of the measurement errors are identically zero, and (as a consequence of the preceding) that the expected values of the observed variables equal the expected values of the corresponding true scores. Symbolically we write

$$E(\underline{u}\,Y) = E(\underline{X}\,v) = \underline{0},\; E(\underline{X}\,\underline{u}^t) = \underline{0}, \text{ and } E(Yv) = 0 \quad , \quad (20)$$

$$E(\underline{u}) = \underline{0} \text{ and } E(v) = 0, \text{ and} \tag{21}$$

$$E(\underline{x}) = E(\underline{X}) \text{ and } E(y) = E(Y) \tag{22}$$

The reader should note that the errors of measurement are permitted to be correlated, e.g., $E(\underline{u}\ v) \neq \underline{0}$, since in many analyses of change it is quite reasonable to expect pretest and posttest errors to be correlated.[1] Now we present the partitioned covariance matrix of the observed y and $\underline{x}$ values:

$$\underline{S}_{yx} = \left|\begin{array}{c|c} s_y^2 & \underline{s}_{yx}^t \\ \hline \underline{s}_{yx} & \underline{S}_{xx} \end{array}\right| . \tag{23}$$

Since the true scores and errors of measurement are uncorrelated $\underline{S}_{yx}$ is the sum of $\underline{S}_{yx}$ and $\underline{S}_{vu}$, and using previous results the covariance structure of the observed scores can be written strictly in terms of the structural parameters:

$$\underline{S}_{yx} = \left[\begin{array}{c|c} s_y^2 + \underline{s}_v^2 & \underline{s}_{YX}^t + \underline{s}_{vu}^t \\ \underline{s}_{YX} + \underline{s}_{vu} & \underline{S}_{XX} + \underline{S}_{uu} \end{array}\right]$$

$$= \left[\begin{array}{c|c} \underline{b}^t \underline{S}_{XX}\, \underline{b} + s_v^2 + s_e^2 & \underline{b}^t \underline{S}_{XX} + \underline{s}_{vu}^t \\ \underline{S}_{XX}\, \underline{b} + \underline{s}_{vu} & \underline{S}_{XX} + \underline{S}_{uu} \end{array}\right] \tag{24}$$

The vector of observed-score regression coefficients defined as

$$\underline{b}' = \underline{S}_{xx}^{-1}\, \underline{s}_{xy} \tag{25}$$

can be expressed in terms of the structural parameters:

$$\underline{b}' = (\underline{S}_{XX} + \underline{S}_{uu})^{-1} (\underline{s}_{YX} + \underline{s}_{vu})$$

$$= [(\underline{S}_{XX} + \underline{S}_{uu})^{-1} \underline{S}_{XX} \underline{b} + (\underline{S}_{XX} + \underline{S}_{uu})^{-1} \underline{s}_{vu}] \quad . \qquad (26)$$

If we let

$$\underline{L} = (\underline{S}_{XX} + \underline{S}_{uu})^{-1} \underline{S}_{XX} \quad \text{and} \quad \underline{m} = (\underline{S}_{XX} + \underline{S}_{uu})^{-1} \underline{s}_{vu} \quad , \qquad (27)$$

then

$$\underline{b}' = \underline{L}\,\underline{b} + \underline{m} \quad . \qquad (28)$$

Thus, the vector of observed-score regression coefficients is seen to be a weighted linear combination of the true-score regression vector and the true, and error covariances. The observed intercept then becomes

$$\underline{b}'_0 = b_0 + (\underline{b} - \underline{b}')\,\underline{\bar{X}} \quad . \qquad (29)$$

$s_{e'}^2$ for Observed Change

The goodness of fit indices of the observed-score regression can also be written in terms of the structural parameters. The first index of fit, the variance of the observed residual, can be derived as

$$s_{e'}^2 = E[\{y-(b'_0 + \underline{b}'^t\underline{x})\}^t \{y-(b'_0 + \underline{b}'\underline{x})\}]$$

$$= E[\{(Y + v) - (b'_0 + \underline{b}'^t (\underline{X} + \underline{u}))\}^t \{(Y + v) - (b'_0 + \underline{b}'^t(\underline{X} + \underline{u}))\}]$$

$$= s_Y^2 + s_v^2 + \underline{b}'^t\underline{S}_{XX}\,\underline{b}' + \underline{b}'^t\,\underline{S}_{uu}\,\underline{b}' - 2\underline{b}'^t\,(\underline{s}_{YX} + s_{vu}) \quad . \qquad (30)$$

Equation 28 can be employed to express $\underline{s}_{e'}^2$ as a function of the parameters of the latent variable distributions:

$$s_{e'}^2 = s_Y^2 + s_v^2 + (\underline{L}\,\underline{b} + \underline{m})^t\,\underline{S}_{XX}(\underline{L}\,\underline{b} + \underline{m}) + (\underline{L}\,\underline{b} + \underline{m})^t\,\underline{S}_{uu}\,(\underline{L}\,\underline{b} + \underline{m})$$

$$- 2\,[(\underline{b} + \underline{m})^t\ (\underline{s}_{YX}\ s_{vu})] \quad . \qquad (31)$$

$R^{2'}$ for Observed Change

Second, the coefficient of multiple determination is given by

$$R^2 \quad \underline{b}'^t \, \underline{S}_{XX} \, \underline{b}' \, / s_y^2 = \underline{b}'^t \, (\underline{S}_{XX} + \underline{S}_{uu}) \underline{b}' \, / (s_Y^2 + s_v^2). \qquad (32)$$

Using equation 28 the coefficient of multiple determination of the observed variables may be written exclusively in terms of the latent variable parameters as

$$R^2 = (\underline{b}^t \underline{L}^t \underline{S}_{XX} \, \underline{L} \, \underline{b} + \underline{b}^t \underline{L}^t \underline{S}_{uu} \, \underline{L} \, \underline{b} + \underline{m}^t \underline{S}_{XX} \, \underline{L} \, \underline{b} + \underline{m}^t \underline{S}_{uu} \, \underline{L} \, \underline{b}$$

$$+ \underline{b}^t \underline{L}^t \underline{S}_{XX} \, \underline{m} + \underline{b}^t \underline{L}^t \, \underline{S}_{uu} \, \underline{m} + \underline{m}^t \, \underline{S}_{XX} \, \underline{m} + \underline{m}^t \, \underline{S}_{uu} \, \underline{m}) \, / \, (\underline{s}_Y^2 + \underline{s}_v^2) \, . \qquad (33)$$

$\underline{S}_{\underline{b}'\underline{b}'}$ for Observed Change

Information about the joint sampling distribution of the observed-score regression coefficients is contained in

$$\underline{S}_{\underline{b}'\underline{b}'} = s_{e'}^2 \cdot \frac{1}{N} \cdot (\underline{S}_{XX} + \underline{S}_{uu})^{-1} \qquad (34)$$

Clearly Equation 31 can be used to write $\underline{S}_{\underline{b}'\underline{b}'}$ as a function of the latent variable parameters:

$$\underline{S}_{\underline{b}'\underline{b}'} = 1/N \cdot \; s_Y^2 + s_v^2 + (\underline{L} \, \underline{b} + \underline{m})^t \, \underline{S}_{XX} \, (\underline{L} \, \underline{b} + \underline{m}) + (\underline{L} \, \underline{b} + \underline{m})^t \, \underline{S}_{uu} \, (\underline{L} \, \underline{b} + \underline{m})$$

$$-2 \, [(\underline{b} + \underline{m})^t \, (\underline{s}_{YX} \, s_{vu})] \quad \cdot \; (\underline{S}_{XX} + \underline{S}_{uu})^{-1} \, . \qquad (35)$$

PARAMETERS AFFECTING BIAS IN OBSERVED-SCORE REGRESSION STATISTICS

Thus far, expressions for the true-score regression parameters ($\underline{b}$, s_e^2, R^2, and $\underline{S}_{bb}$) and the observed-score regression parameters ($\underline{b}'$, $s_{e'}^2$, $R^{2'}$, and $\underline{S}_{\underline{b}'\underline{b}'}$) have been derived exclusively in terms of the parameters of the joint distributions of the true scores, $\underline{X}$ and Y. In the following sections, we compare the parametric expressions for pairs of corresponding true-score and observed-score regression statistics. This process enables us to state some new analytic results demonstrating how the bias the in observed-score regression estimators is affected by the distributions of the true and error components of the observed scores. In many cases, however, simple general statements cannot be made without making strong assumptions because of the mathematical complexities. Even the general expressions provide insights into the biasing effects of errors of measurement and enable investigators to estimate a priori the degree of bias that is likely to be found in most studies of change.

Parameters Affecting Bias in $s_{e'}^2$

To proceed, Equation 14 for s_e^2 and Equation 31 for $s_{e'}^2$ are segregated into three corresponding units based on their comparable structure. These are labeled A, B and C for s_e^2 and A', B', and C' for $s_{e'}^2$.

	s_e^2		$s_{e'}^2$
A:	s_Y^2	A':	$s_Y^2 + s_v^2$
B:	$\underline{b}^t \underline{S}_{XX}\, \underline{b}$	B':	$(\underline{L}\,\underline{b} + \underline{m})^t\, \underline{S}_{XX}\, (\underline{L}\,\underline{b} + \underline{m}) + (\underline{L}\,\underline{b} + \underline{m})^t\, \underline{S}_{uu}\, (\underline{L}\,\underline{b} + \underline{m})$
C:	$-2\underline{b}^t \underline{s}_{YX}$	C':	$-2[(\underline{L}\,\underline{b} + \underline{m})^t\, (\underline{s}_{YX} + \underline{s}_{vu})$

where from Equation 27

$$\underline{L} = (\underline{S}_{XX} + \underline{S}_{uu})^{-1}\, \underline{S}_{XX} \text{ and } \underline{m} = (\underline{S}_{XX} + \underline{S}_{uu})^{-1}\, \underline{s}_{vu}$$

It can be seen immediately by comparing A and A' that the observed-score residual will exceed the true-score residual when s_v^2 is greater than zero. The discrepancy will increase as the magnitude of

s_v^2 increases. Since power, or one minus the probability of Type II error, is an inverse function of $s_{e'}^2$, it is clear that measurement error in the criterion reduces the power of observed-score vis a vis true-score regression tests.

As $\underline{s}_{vu}$ increases, C' decreases and B' increases relative to C and B (holding all other terms constant). The effect of $\underline{s}_{vu}$ on the observed residual depends upon the magnitude of $\underline{S}_{XX}$ and $\underline{S}_{uu}$ relative to $\underline{s}_{YX}$. In general, positive covariances among the criterion and predictor measurement errors will reduce the bias in $s_{e'}^2$ as an estimator of s_e^2. Negative covariances, however, will tend to increase the bias. It seems impossible to make a general statement about the absolute difference between $s_{e'}^2$ and s_e^2 as a function of $\underline{s}_{vu}$. The actual degree of bias will vary with the size and pattern of incorrelation among the $\underline{X}$ and the $\underline{u}$. It can be concluded, however, that in general bias will increase as $\underline{s}_{vu}$ decreases. This is probably a fortunate result for investigations of change, because error covariances among pre- and posttest measurements will be positive in most circumstances.

The effect of $\underline{S}_{uu}$ on $s_{e'}^2$ is difficult to assess since it appears in $\underline{L}$, $\underline{m}$, and B'. When $\underline{S}_{uu}$ is diagonal and $\underline{s}_{vu} = \underline{0}$, as measurement error variances increase, the larger $s_{e'}^2$ will be relative to s_e^2. The effect of $\underline{S}_{uu}$ on the bias in $s_{e'}^2$ can not be ascertained for the general situation in which the errors of measurement may be positively or negatively interrelated.

Finally, evaluation of B' and C' reveals that the bias in $s_{e'}^2$ will be reduced as $\underline{S}_{XX}$ dominates $\underline{S}_{uu}$ in $\underline{S}_{xx}$. As S_{XX} approaches

S_{uu} in value, $s_{e'}^2$ approaches s_e^2. However, the pattern of relations among the $\underline{X}$ and among the $\underline{u}$ can nullify this.

In summary, the bias in $s_{e'}^2$ will increase as a function of the variance of the measurement errors in the dependent and independent variables and the covariances of the measurement errors in the predictor variables. Negatively correlated criterion and predictor measurement errors tend to increase the bias. In most analyses of educational change, measurement error will reduce the power of statistical tests, decrease the precision of parameter estimation, and increase the probability of inferential errors of the second kind.

Parameters Affecting Bias in $R^{2'}$

Reference to Equations 15 and 33 indicate that the following segmentation can be made:

R^2	$R^{2'}$
A: $1/(s_Y^2)$	A': $1/(s_Y^2 + s_V^2)$
B: $\underline{b}^t \underline{S}_{XX} \underline{b}$	B': $\underline{b}^t \underline{L}^t \underline{S}_{XX} \underline{L}\, \underline{b} + \underline{b}^t \underline{L}\, \underline{S}_{uu} \underline{L}\, \underline{b}$
	$+ \underline{m}^t \underline{S}_{XX} \underline{L}\, \underline{b} + \underline{m}^t \underline{S}_{uu} \underline{L}\, \underline{b}$
	$+ \underline{b}^t \underline{L}^t \underline{S}_{XX} \underline{m} + \underline{b}^t \underline{L}^t \underline{S}_{uu} \underline{m}$
	$+ \underline{m}^t \underline{S}_{XX} \underline{m} + \underline{m}^t \underline{S}_{uu} \underline{m}$

Inspection of A' indicates that errors of measurement in the criterion variable negatively bias the estimation of the squared multiple correlation. As the unreliability of y increases, the bias (underestimation) of $R^{2'}$ grows. Thus, both goodness of fit parameters

($s_{e'}^2$ and $R^{2'}$) are negatively biased by errors of measurement in the criterion.

The effect of $\underline{s}_{uv}$ on bias in the squared multiple correlation is similar to its effect on the regression residual as demonstrated in the preceeding section. Negative covariances among the criterion and predictor measurement errors will increase the bias in $\hat{R}^{2'}$. Positive covariances will tend to decrease the bias. The actual amount of bias ($= R^{2'} - R^2$) is a complex function of $\underline{S}_{XX}$ and $\underline{S}_{uu}$ as well as $\underline{s}_{vu}$. General statements do not appear possible.

Measurement errors in the predictors affect the bias in $R^{2'}$ in a complex way. The role of $\underline{S}_{uu}$ in $\underline{L}$ increases bias as long as the covariances are positive. On the other hand, the separate terms involving $\underline{S}_{uu}$ in B' tend to decrease bias when the error covariances are greater than zero. The total effect of $\underline{S}_{uu}$ on bias will depend, therefore, on the actual values of $\underline{S}_{uu}$ and $\underline{S}_{XX}$. Working through a series of examples indicates that in most analyses of educational change, the overall effect of predictor measurement errors will be to increase the bias in the estimate of the squared multiple correlation. This assumes that most error covariances are positive and small in size. The degree of bias decreases as $\underline{S}_{XX}$ dominates $\underline{S}_{uu}$. In conclusion, observed-score regression analyses of change on the average will underestimate the goodness of fit of the model in most applications. "On the average" does not mean "always" so investigators should be cautious in assuming that the squared multiple correlation estimate has been attenuated.

Parameters Affecting Bias in $\underline{S}_{\underline{b}'\underline{b}'}$

Reference to Equations 16 and 34 reveals that the following structural comparisons can be made for $\underline{S}_{bb}$ and $\underline{S}_{\underline{b}'\underline{b}'}$:

$\underline{S}_{bb}$	$\underline{S}_{\underline{b}'\underline{b}'}$
A: $s_e{}^2$	A': $s_{e'}{}^2$
B: $(1/N)\underline{S}_{XX}{}^{-1}$	B': $(1/N)(\underline{S}_{XX} + \underline{S}_{uu})^{-1}$

A and A' indicate that the factors which affect $s_{e'}{}^2$ will influence $\underline{S}_{\underline{b}'\underline{b}'}$ in the same ways. Thus, the estimates of the standard errors will be inflated by posttest measurement errors and negative criterion-predictor measurement error covariances. The effect of $\underline{S}_{uu}$ is difficult to assess for the general case. The elements of $\underline{S}_{\underline{b}'\underline{b}'}$ tend to increase as the variances and covariances in $\underline{S}_{uu}$ increase. It is the pattern of elements in $\underline{S}_{uu}$, however, which determines the extent of bias in $\underline{S}_{\underline{b}'\underline{b}'}$ generally.

The effect of the patterns of interrelations among the true and error components on bias in $\underline{S}_{\underline{b}'\underline{b}'}$ is most apparent in segments B and B'. For the situation in which $\underline{S}_{uu}$ is diagonal and small relative to $\underline{S}_{XX}$, predictor errors of measurement make the sampling distribution estimates too large. When $\underline{S}_{uu}$ is nondiagonal containing both positive and negative covariances which approximate the elements of $\underline{S}_{XX}$ in value, a general result concerning the bias in $\underline{S}_{\underline{b}'\underline{b}'}$ cannot be derived. In conclusion, measurement errors tend to make estimates of the regression coefficients less precise than they would be if perfectly reliable variables were used. The degree of bias is a joint function of $s_v{}^2$, $\underline{s}_{uv}$, $\underline{S}_{uu}$, and $\underline{S}_{XX}$. Statements that apply across all conditions and patterns of relationshp can not be made, however.

CONCLUSION

In this chapter general matrix expressions for the true-score regression parameters have been given. The observed-score regression parameters were expressed as functions of the true-score regression parameters and the true and error covariance structures. The parameters affecting bias in the observed-score regression statistics were evaluated by comparing the expressions for the observed- and true-score coefficients. Specifically, the biasing effects of s_v^2, $\underline{s}_{uv}$, $\underline{S}_{uu}$, and $\underline{S}_{XX}$ on s_e^{2}, R^2, and $\underline{S}_{b'b'}$ were explored. Some unequivocal statements could be made, e.g., bias increases in all observed-score estimators as a direct function of s_v^2. By making strong assumptions about the error structure, viz. that $\underline{s}_{uv}$ equals $\underline{0}$ and $\underline{S}_{uu}$ is diagonal, other general statements could be made, e.g., bias increases as $\underline{S}_{uu}$ increases. However, it was not possible to draw unqualified general conclusions about the parametric determinants of bias. Much insight into the biasing effects of measurement error has been gained by examination of the expressions derived in this chapter. In the next chapter these formulas will permit development of an algorithm that can be used in studies of change to assess the potential bias caused by the unreliability of measures.

CHAPTER VI

AN ALGORITHM FOR ASSESSING BIAS IN PLANNED STUDIES OF CHANGE

INTRODUCTION

The purpose of this chapter is to develop a method for investigators to easily assess the possible impact of measurement error on statistical analyses of change. Using the results of the preceeding chapters, especially those of Chapter V, an algorithm is developed which takes as input estimates of the parameter values of the structural relations among the latent variables (which the investigator thinks are close to the true values a priori) and outputs the expected values of the corresponding observed-score regression parameters for a prespecified sample size. The logic of the algorithm is explained and illustrated with a simple example of the effects of external locus of control orientation on change in science achievement.

As part of this research program, the algorithm was implemented in the form of a FORTRAN computer program, which can be easily installed in most software libraries. The program enables researchers to input a series of estimates of the true-score parameter values and obtain expected values of the corresponding observed-score regressions. In the final section of the chapter, a comprehensive application of the computer program is presented.

Use of the program will enable investigators to become aware of the ways in which measurement error may bias regression analyses of change. Making this evaluation before data collection is completely analogous to carrying out a power analysis. The results of the assessment may lead the investigator to modify data collection plans. For example, the program may reveal that the reliability of the pretest must be increased if accurate inferences are to be possible. The assessment may indicate that bias can not be avoided easily and prompt the investigator to gather the data in such a way as to make the use of attenuation-correction methods or multiple indicator (LISREL) models possible. Also, as with power analysis, the program can be used post hoc to determine the degree of caution one should have when interpreting the results of the regression analyses of observed scores. In many situations, like the one described in the example in this chapter, it will be concluded that the possible bias in the observed-score regression estimators was so great that any inferences must be regarded as completely suspect.

THE ALGORITHM

The algorithm requires information about the structural parameters and the variance structure of the true and error components as input. Specifically, hypothesized or likely values of $\underline{S}_{XX}$, $\underline{S}_{uu}$, $\underline{s}_{Y}^{2}$, $\underline{s}_{vu}$, s_{v}^{2}, and $\underline{b}$ are necessary. When in this chapter, two simplifying assumptions are made, viz., that $\underline{S}_{uu}$ is diagonal and $\underline{s}_{uv} = \underline{0}$, information about the reliabilities of the observed predictor variables can be used instead of error variances. The algorithm, however, is developed in a general form that will accomodate any measurement error variance structure. In the following we let q equal the number of predictor variables in the model plus one.

The algorithm first computes the important true-score regression statistics, R^2, s_e^2, and $\underline{S}_{bb}$ as follows:

$$R^2 = (\underline{b}^t \underline{S}_{XX} \underline{b}) / s_Y^2 \quad , \tag{1}$$

$$s_e^2 = [N/(N-q)] \, s_Y^2 \, (1-R^2) \quad , \tag{2}$$

$$\underline{S}_{bb} = s_e^2 \, [N/(N-q)] \, \underline{S}_{XX}^{-1} \quad , \tag{3}$$

$$\underline{s}_{YX} = \underline{S}_{XX} \, \underline{b} \quad . \tag{4}$$

If one had information about $\underline{s}_{YX}$ instead of $\underline{b}$, Equation 14 could be solved first and then Equations 1-3. These equations have been derived in more complicated forms in the preceeding chapters. For ease of application, they are presented in their simplest or most easily calculable form here. The t-tests associated with the hypothesis that the regression coefficient equals zero in the population are determined next:

$$t_{b_1} = b_1 / s_{b_1} \, , \; t_{b_2} = b_2 / s_{b_2} \, , \; \ldots \; , \tag{5}$$

where the probability associated with each t is a function of N-q degrees of freedom.

In the next stage of the algorithm, the variance structure of the observed $\underline{x}$ and y scores are derived:

$$\underline{S}_{xx} = \underline{S}_{XX} + \underline{S}_{uu} \quad , \tag{6}$$

$$\underline{s}_{yx} = \underline{s}_{YX} + \underline{s}_{vu} \quad , \tag{7}$$

$$s_y^2 = s_Y^2 + s_v^2 \quad . \tag{8}$$

Standard regression formulas are then applied to the observed-score covariance matrix to derive estimates of the observed regression parameters. Observed-score parameters, corresponding to the true-score parameters given by Equations 1-3 and 5, are found as follows:

$$R^{2'} = (\underline{b}'^t \underline{S}_{xx} \underline{b}') / s_y^2 \quad , \tag{9}$$

$$s_{e'}^2 = [N/(N-q)] \, s_y^2 \, (1-R^{2'}) \quad , \tag{10}$$

$$\underline{S}_{\underline{b}'\underline{b}'} = s_{e'}^2 \, [1/(N-q)] \, \underline{S}_{xx}^{-1} \tag{11}$$

$$t_{b'_1} = b'_1 / s_{b'_1} , \; t_{b'_2} = b'_2 / s_{b'_2} , \; . \; . \; . \; . \tag{12}$$

The estimates and significance test results for the true-score and observed-score distributions can be compared with ease and the potential for bias and incorrect inferences assessed.

AN EXAMPLE

To illustrate the value and use of the algorithm a brief example is presented. Consider an investigation which seeks to test the hypothesis that external locus of control orientation exerts a negative effect on true change in science achievement. The variables are posttest science achievement (Y), pretest science achievement (X_1), and external locus of control (X_2). It is assumed that the correlation between true pretest achievement and true external locus of control is .74, and that the effect of locus of control on true change is -.230. The reliabilities of the science pretest and external locus of control scales are assumed to be .769 and .951, respectively. Complete input information includes:

$$\underline{S}_{XX} = \begin{bmatrix} 1.00 & .74 \\ .74 & 1.00 \end{bmatrix} \quad ,$$

$$\underline{S}_{uu} = \begin{bmatrix} .300 & .000 \\ .000 & .051 \end{bmatrix},$$

$$\underline{s}_{vu} = \begin{bmatrix} .000 \\ .000 \end{bmatrix},$$

$$\underline{b} = \begin{bmatrix} 1.028 \\ -.230 \end{bmatrix},$$

$s_Y^2 = 1.0$, $N = 200$, and $q = 3$.

The true-score regression parameters are found using Equations 1 through 5:

$$R^2 = .860$$

$$s_e^2 = .141$$

$$\underline{S}_{\underline{bb}} = \begin{bmatrix} .002 & -.001 \\ -.001 & .002 \end{bmatrix}$$

$$\underline{S}_{YX} = \begin{bmatrix} .86 \\ .53 \end{bmatrix}$$

$$t_{b_1} = 23.00 , \quad t_{b_2} = -5.15 .$$

With a sample of 200 observations the regression coefficients would be extremely well estimated. It is clear that a substantial portion of the true posttest achievement variance can be explained on the basis of the true pretest and locus of control ($R^2 = .86$). The inference that the effect of external locus of control is negative would be strongly supported [t_{b_2} (197) = −5.15, $p < .001$].

Having derived the true-score regression coefficients for the hypothesized true-score distribution, Equations 9-12 are used to obtain the corresponding observed-score coefficients. First, the observed correlation matrix is calculated

$$\underline{S}_{xx} = \begin{bmatrix} 1.00 & .60 \\ .60 & 1.00 \end{bmatrix}$$

and then the observed predictor-criterion covariance vector:

$$\underline{s}_{yx} = \begin{bmatrix} .70 \\ .50 \end{bmatrix}$$

The regression estimates are then easily computed:

$$\underline{b}' = \begin{bmatrix} .63 \\ .13 \end{bmatrix}$$

$$\begin{matrix} R^{2\prime} = \\ s_{e'}^{2} = \end{matrix} \quad \begin{bmatrix} .50 \\ .50 \end{bmatrix}$$

$$\underline{S}_{\underline{b}'\underline{b}'} = \begin{bmatrix} .004 & -.002 \\ -.002 & .004 \end{bmatrix}$$

$$t_{b'_1} = 9.96, \quad t_{b'_2} = 2.06 \quad .$$

The observed-score regression weight for the pretest is attenuated, as would be expected from the unreliability of the science pretest. Much to our hypothetical investigator's chagrin, the observed-score regression estimate of the effect of external locus of control is (significantly) positive! Thus, inferences about the effects of external locus of control on change in science achievement would be completely erroneous if the fallibility of the measures was not recognized. Use of the algorithm, however, alerted our hypothetical researcher to the potential danger, thus enabling him or her to take corrective actions prior to data collection or to be appropriately cautious in interpreting the results if the study had already been completed. It is also worth noting that the algorithm showed that on the average the observed-score model would evidence less goodness of fit and lower power.

FORTRAN COMPUTER PROGRAM

The algorithm described above was implemented as a FORTRAN program as one part of the overall research effort. It is written in standard FORTRAN and, although it was run on the WATFIV compiler at New York University, the program can be installed without much effort on any computer system. Furthermore, the structure allows it to be modified easily to handle more general problems, e.g., more than two predictors. The program is designed to be optimally useful to investigators who are planning a study of change (but have not yet begun data collection).

As the flow charts in Figures 1 and 2 show, the program follows the structure of the algorithm presented above very closely. A source

Figure 1

MAIN PROGRAM

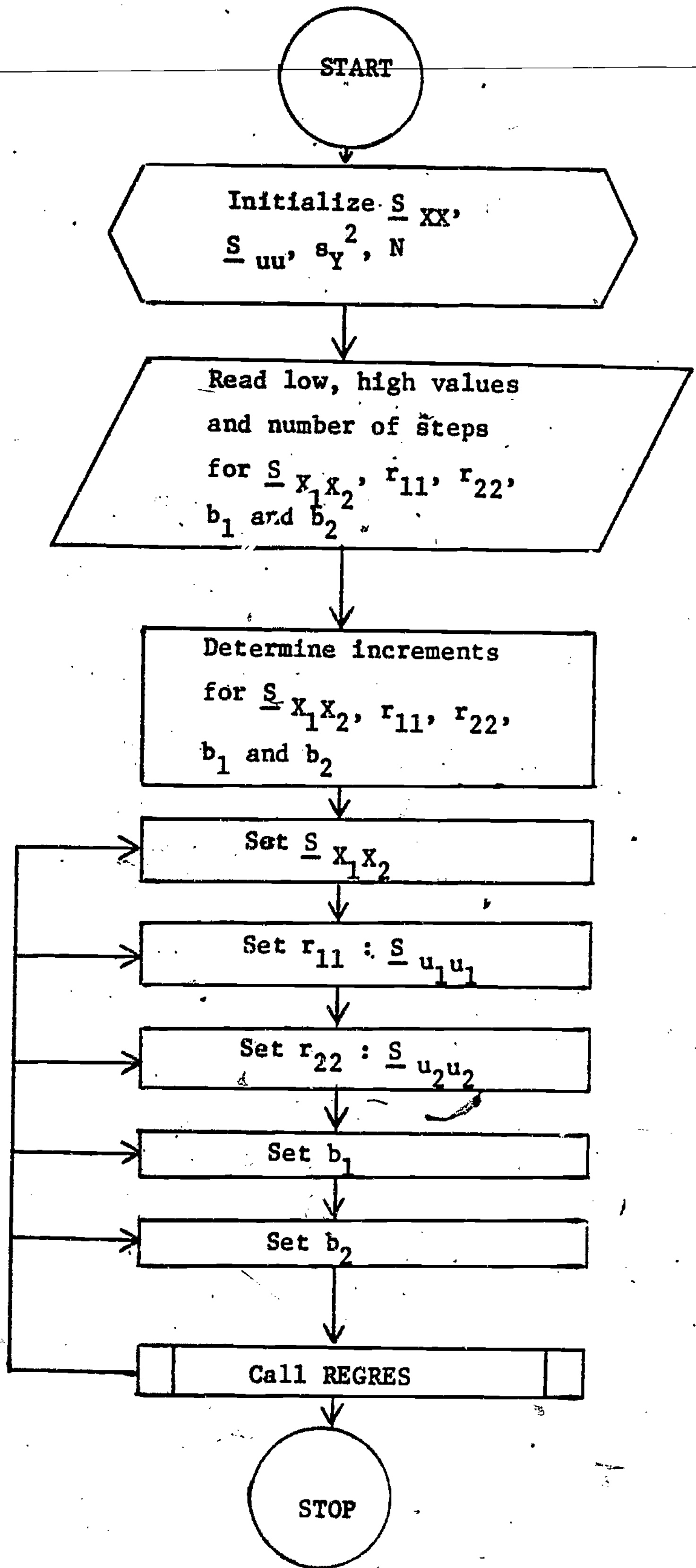
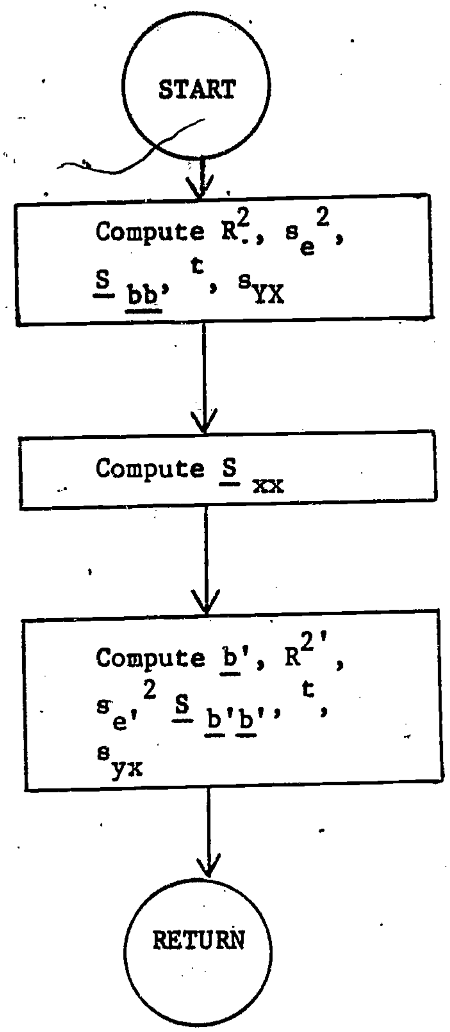


Figure 2

REGRES SUBROUTINE



Other Subroutines:

MVMAT:	Computes matrix-vector product.
MINV:	Computes matrix inverse.
QUADF:	Computes quadratic form.

listing of the program appears in the appendix to this chapter. Input to the program consists of information about the variance structure of the true scores, the true-score regression coefficients, and the reliabilities of the observed predictors. The program allows calculation of several sets of parameter values, thus the input specifies a range of values and the number of estimates to be calculated within that range. In the version of the program illustrated in this chapter, two predictors, X_1, and X_2, are permitted. Let n_1 refer to the number of possible values of $s_{X_1X_2}$ that are specified by the input, n_2 to the number of levels of r_{11} ($s_{u_1u_2}$), n_3 to the number of levels of r_{22} ($s_{u_2u_2}$), n_4 to the number of values of b_1, and n_5 to the number of values of b_2. Then one run of the program produces $n_1 \times n_2 \times n_3 \times n_4 \times n_5$ combinations of solutions.

This iterative feature was included since researchers will often have little confidence in their specific a priori expectations about true-score regression parameters. Frequently, however, a range of possible parameter values can be stated with some confidence. The program assesses the potential biases for all combinations of suspected parameter values in a single run. Although the program is formulated in terms of covariances, it will be used most often with standardized estimates. Hence, all illustrations below are given in terms of correlations and standard regression weights.

Once the main program has converted the reliabilities into error variances and calculated the increments in parameter values which cover the prespecified range from lowest to highest values, a subroutine which performs the major computations is called for each combination of parameter values. On each call, the true-score regression parameters R^2, s_e^2, $\underline{S}_{\underline{bb}}$, t_{b_1}, t_{b_2}, and $\underline{s}_{YX}$ are calculated first. Then the covariance matrix of the observed scores, $\underline{S}_{xx}$, is determined. Finally, the values of the observed-score regression parameters, $\underline{b}'$, $R^{2'}$, $s_{e'}^2$, $\underline{S}_{\underline{b}'\underline{b}'}$, $t_{b'_1}$, $t_{b'_2}$, and $\underline{s}_{yx}$, are obtained and printed. The program terminates after the final call to the subroutine.

AN APPLICATION

To illustrate the use of the program, it was applied to the following problem. An investigator was planning a study of change in which it was anticipated that b_2 could range between −.4 and .4, b_1 between .1 and .7, r_{22} between .7 and .9, r_{11} between .6 and .9, and $r_{X_1X_2}$ between −.4 and .4. Within this set of conditions, what degree of bias could be anticipated in the observed-score regression coefficients as estimators of the true-score regression parameters? The program evaluated 2 x 2 x 3 x 2 x 3 = 72 combinations of conditions and printed the results in Table 1.

Insert Table 1 about here

Comparison of the true-score squared multiple correlation (column R2T) with the observed-score squared multiple correlation (column R2O) indicates that little bias should be expected. As long as the reliabilities are high, $R^{2'}$ almost equals R^2. When the reliabilities drop, $R^{2'}$ underestimates R^2 by .1 to .2. Consistent with this result is the comparison of s_e^2 (VET) and $s_{e'}^2$ (VEO), which indicates that $s_{e'}^2$ is inflated only for combinations of low reliabilities

(r_{11} = .6, r_{22} = .7). The bias in neither goodness of fit index seems large enough to cause the investigator much concern. Observed-score results concerning the adequacy of the model should be reasonably close to the true-score parameters on the average.

The major statistics of interest in the study of change are b_2 and $b_{2'}$. The columns headed TB2 and OB2 (b_2 and b'_2, respectively) show that there is some bias. The degree of bias is as great as −.2 [−.2 − (−.4) and .2 −.4] in absolute terms and 50% on a relative basis (−.2/−.4 and .2/.4). For all combinations except three, the observed-score coefficient falls within the −.4 to .4 range. In the null case, that is, where b_2 = 0, the observed-score bias is quite small across all parameter combinations. Examination of the t-ratios for b_2 and b'_2 (columns T-TB2 and T-OB2, respectively) reveals that the true-score and observed-score results are perfectly consistent. That is, there are no combinations of nonnull conditions for which t_{b_2} is significant but $t_{b'_2}$ is not, and *vice versa*. When the true-score regression weight equals 0, there are no instances where T-OB2 leads to rejection of the hypothesis that the observed-score regression coefficient equals 0.

Although some power will be lost as a result of unreliability, our hypothetical researcher learns from the output of the program that measurement error will not lead to incorrect inferences about the effect of X_2 on true change (i.e., about b_2) on the average. Of course, this was the major concern that motivated the preliminary analysis of potential bias due to unreliability. In this situation the investigator may well decide to proceed with collecting data on x_1, x_2, and y and subsequently performing a regression analysis of the observed scores.

The advisability of this course of action depends upon how closely the hypothesized parameter values approximate the actual values in the population from which the researcher will draw the sample. In many other circumstances the opposite course will be decided after examining the output of the program, e.g., in situations like the science achievement example presented earlier.

CONCLUSION

In this chapter we have described the development of a FORTRAN program which is based upon an algorithm that expresses both the true-score and observed-score regression parameters as functions of the variance structures of the true and error components. Input to the program consists of information about the covariances among the true predictors, the reliabilites of the observed predictors, and the true-score regression coefficients. The program outputs values of the true-score regression parameters and those of the corresponding observed-score regression parameters. Comparison of the two sets of parameter values allows one to assess the degree of bias likely to occur in observed-score regression coefficients as estimators of their true-score counterparts. Preliminary evaluation of potential errors of inference due to measurement error allows the investigator to redesign the research plan or select new measures. Use of the algorithm and program is strongly recommended, since it can improve the quality of research on the determinants of change and prevent erroneous inferences.

A DEMONSTRATION OF THE RELATIONSHIP BETWEEN PARAMETERS OF TRUE-SCORE AND OBSERVED-SCORE DISTRIBUTIONS
FOR SELECTED VALUES OF RT12, R11, R22, TB1 AND TB2

Table 1

RT12	R11	TB1	OB1	STB1	SOB1	T-TB1	T-OB1	R22	TB2	OB2	STB2	SOB2	T-TB2	T-OB2	R2T	R2O	VET	VEO	N
-0.4	0.6	0.100	0.096	0.099	0.078	1.010	1.223	0.7	-0.400	-0.281	0.099	0.081	-4.042	-3.454	0.2	0.1	0.8	0.9	100
-0.4	0.6	0.100	0.062	0.110	0.085	0.907	0.735	0.7	0.000	-0.011	0.110	0.088	0.000	-0.120	0.0	0.0	1.0	1.0	100
-0.4	0.6	0.100	0.029	0.103	0.081	0.972	0.353	0.7	0.400	0.260	0.103	0.084	3.889	3.093	0.1	0.1	0.9	0.9	100
-0.4	0.6	0.400	0.283	0.082	0.069	4.860	4.073	0.7	-0.400	-0.313	0.082	0.072	-4.860	-4.344	0.4	0.3	0.6	0.7	100
-0.4	0.6	0.400	0.249	0.102	0.080	3.940	3.098	0.7	0.000	-0.042	0.102	0.083	0.000	-0.507	0.2	0.1	0.9	0.9	100
-0.4	0.6	0.400	0.215	0.100	0.080	4.017	2.680	0.7	0.400	0.228	0.100	0.083	4.017	2.738	0.2	0.1	0.8	0.9	100
-0.4	0.6	0.700	0.469	0.039	0.051	17.801	9.177	0.7	-0.400	-0.345	0.039	0.053	-10.172	-6.491	0.9	0.6	0.1	0.4	100
-0.4	0.6	0.700	0.436	0.079	0.070	8.840	6.242	0.7	0.000	-0.074	0.079	0.072	0.000	-1.021	0.5	0.3	0.5	0.7	100
-0.4	0.6	0.700	0.402	0.084	0.074	8.340	5.420	0.7	0.400	0.197	0.084	0.077	4.766	2.557	0.4	0.2	0.6	0.8	100
-0.4	0.6	0.100	0.073	0.099	0.078	1.010	0.937	0.9	-0.400	-0.370	0.099	0.092	-4.042	-4.040	0.2	0.2	0.8	0.8	100
-0.4	0.6	0.100	0.061	0.110	0.086	0.907	0.716	0.9	0.000	-0.014	0.110	0.101	0.000	-0.138	0.0	0.0	1.0	1.0	100
-0.4	0.6	0.100	0.050	0.103	0.081	0.972	0.619	0.9	0.400	0.342	0.103	0.095	3.889	3.603	0.1	0.1	0.9	0.9	100
-0.4	0.6	0.400	0.257	0.082	0.068	4.860	3.779	0.9	-0.400	-0.411	0.082	0.080	-4.860	-5.142	0.4	0.4	0.6	0.6	100
-0.4	0.6	0.400	0.246	0.102	0.081	3.940	3.022	0.9	0.000	-0.056	0.102	0.096	0.000	-0.582	0.2	0.1	0.9	0.9	100
-0.4	0.6	0.400	0.234	0.100	0.080	4.017	2.916	0.9	0.400	0.300	0.100	0.094	4.017	3.179	0.2	0.1	0.8	0.9	100
-0.4	0.6	0.700	0.441	0.039	0.048	17.801	9.180	0.9	-0.400	-0.453	0.039	0.057	-10.172	-8.013	0.9	0.7	0.1	0.3	100
-0.4	0.6	0.700	0.430	0.079	0.070	8.840	6.097	0.9	0.000	-0.097	0.079	0.083	0.000	-1.173	0.5	0.3	0.5	0.7	100
-0.4	0.6	0.700	0.418	0.084	0.074	8.340	5.642	0.9	0.400	0.259	0.084	0.087	4.766	2.965	0.4	0.3	0.6	0.8	100
-0.4	0.9	0.100	0.129	0.099	0.071	1.010	1.419	0.7	-0.400	-0.272	0.099	0.082	-4.042	-3.310	0.2	0.2	0.8	0.9	100
-0.4	0.9	0.100	0.083	0.110	0.098	0.907	0.851	0.7	0.000	-0.005	0.110	0.089	0.000	-0.052	0.0	0.0	1.0	1.0	100
-0.4	0.9	0.100	0.038	0.103	0.094	0.972	0.409	0.7	0.400	0.263	0.103	0.085	3.889	3.088	0.1	0.1	0.9	0.9	100
-0.4	0.9	0.400	0.379	0.082	0.078	4.860	4.850	0.7	-0.400	-0.206	0.082	0.071	-4.860	-4.041	0.4	0.4	0.6	0.6	100
-0.4	0.9	0.400	0.334	0.102	0.091	3.940	3.649	0.7	0.000	-0.019	0.102	0.083	0.000	-0.224	0.2	0.1	0.9	0.9	100
-0.4	0.9	0.400	0.289	0.100	0.092	4.017	3.142	0.7	0.400	0.249	0.100	0.083	4.017	2.985	0.2	0.1	0.8	0.9	100
-0.4	0.9	0.700	0.629	0.039	0.050	17.801	12.655	0.7	-0.400	-0.300	0.039	0.045	-10.172	-6.648	0.9	0.7	0.1	0.3	100
-0.4	0.9	0.700	0.584	0.079	0.075	8.840	7.777	0.7	0.000	-0.032	0.079	0.068	0.000	-0.477	0.5	0.4	0.5	0.6	100

-0.4	0.9	0.700	0.539	0.084	0.081	8.340	6.636	0.7	0.400	0.235	0.084	0.074	4.766	3.187	0.4	0.3	0.6	0.7	100
-0.4	0.9	0.100	0.078	0.099	0.090	1.010	1.090	0.9	-0.400	-0.361	0.099	0.093	-4.042	-3.883	0.2	0.2	0.8	0.8	100
-0.4	0.9	0.100	0.083	0.110	0.099	0.907	0.033	0.9	0.000	-0.006	0.110	0.102	0.000	-0.060	0.0	0.0	1.0	1.0	100
-0.4	0.9	0.100	0.067	0.103	0.094	0.972	0.720	0.9	0.400	0.348	0.103	0.096	3.889	3.613	0.1	0.1	0.9	0.9	100
-0.4	0.9	0.400	0.347	0.082	0.077	4.860	4.510	0.9	-0.400	-0.379	0.082	0.079	-4.860	-4.786	0.4	0.4	0.6	0.6	100
-0.4	0.9	0.400	0.332	0.102	0.093	3.940	3.572	0.9	0.000	-0.025	0.102	0.096	0.000	-0.258	0.2	0.1	0.9	0.9	100
-0.4	0.9	0.400	0.316	0.100	0.092	4.017	3.442	0.9	0.400	0.330	0.100	0.095	4.017	3.489	0.2	0.2	0.8	0.9	100
-0.4	0.9	0.700	0.596	0.039	0.047	17.801	12.793	0.9	-0.400	-0.397	0.039	0.048	-10.172	-8.294	0.9	0.8	0.1	0.2	100
-0.4	0.9	0.700	0.580	0.079	0.076	8.840	7.615	0.9	0.000	-0.043	0.079	0.078	0.000	-0.549	0.5	0.4	0.5	0.6	100
-0.4	0.9	0.700	0.565	0.084	0.081	8.340	6.971	0.9	0.400	0.311	0.084	0.083	4.766	3.734	0.4	0.3	0.6	0.7	100
0.4	0.6	0.100	0.029	0.103	0.081	0.972	0.353	0.7	-0.400	-0.260	0.103	0.084	-3.889	-3.073	0.1	0.1	0.9	0.9	100
0.4	0.6	0.100	0.062	0.110	0.085	0.907	0.735	0.7	0.000	0.011	0.110	0.088	0.000	0.120	0.0	0.0	1.0	1.0	100
0.4	0.6	0.100	0.096	0.099	0.078	1.010	1.223	0.7	0.400	0.201	0.099	0.081	4.042	3.454	0.2	0.1	0.8	0.9	100
0.4	0.6	0.400	0.215	0.100	0.080	4.017	2.600	0.7	-0.400	-0.220	0.100	0.083	-4.017	-2.738	0.2	0.1	0.8	0.9	100
0.4	0.6	0.400	0.249	0.102	0.080	3.940	3.070	0.7	0.000	0.042	0.102	0.083	0.000	0.507	0.2	0.1	0.9	0.9	100
0.4	0.6	0.400	0.283	0.082	0.069	4.860	4.073	0.7	0.400	0.313	0.082	0.072	4.860	4.344	0.4	0.3	0.6	0.7	100
0.4	0.6	0.700	0.402	0.084	0.074	8.340	5.420	0.7	-0.400	-0.197	0.084	0.077	-4.766	-2.557	0.4	0.2	0.6	0.8	100
0.4	0.6	0.700	0.436	0.079	0.070	8.840	6.242	0.7	0.000	0.074	0.079	0.072	0.000	1.021	0.5	0.3	0.5	0.7	100
0.4	0.6	0.700	0.469	0.039	0.051	17.801	9.177	0.7	0.400	0.345	0.039	0.053	10.172	6.491	0.9	0.6	0.1	0.4	100
0.4	0.6	0.100	0.050	0.103	0.081	0.972	0.619	0.9	-0.400	-0.342	0.103	0.095	-3.889	-3.603	0.1	0.1	0.9	0.9	100
0.4	0.6	0.400	0.061	0.110	0.086	0.907	0.716	0.9	0.000	0.014	0.110	0.101	0.000	0.138	0.0	0.0	1.0	1.0	100
0.4	0.6	0.100	0.073	0.099	0.078	1.010	0.937	0.9	0.400	0.370	0.099	0.092	4.042	4.040	0.2	0.2	0.8	0.8	100
0.4	0.6	0.400	0.234	0.100	0.080	4.017	2.916	0.9	-0.400	-0.300	0.100	0.094	-4.017	-3.179	0.2	0.1	0.8	0.9	100
0.4	0.6	0.400	0.246	0.102	0.081	3.940	3.022	0.9	0.000	0.056	0.102	0.096	0.000	0.582	0.2	0.1	0.9	0.9	100
0.4	0.6	0.400	0.257	0.082	0.068	4.860	3.779	0.9	0.400	0.411	0.082	0.080	4.860	5.142	0.4	0.4	0.6	0.6	100
0.4	0.6	0.700	0.418	0.084	0.074	8.340	5.642	0.9	-0.400	-0.259	0.084	0.087	-4.766	-2.965	0.4	0.3	0.6	0.8	100
0.4	0.6	0.700	0.430	0.079	0.070	8.848	6.097	0.9	0.000	0.097	0.079	0.083	0.000	1.173	0.5	0.3	0.5	0.7	100
0.4	0.6	0.700	0.441	0.039	0.048	17.801	9.100	0.9	0.400	0.453	0.039	0.057	10.172	8.013	0.9	0.7	0.1	0.3	100
0.4	0.9	0.100	0.038	0.103	0.094	0.972	0.409	0.7	-0.400	-0.263	0.103	0.085	-3.889	-3.088	0.1	0.1	0.9	0.9	100
0.4	0.9	0.100	0.083	0.110	0.098	0.907	0.851	0.7	0.000	0.005	0.110	0.089	0.000	0.052	0.0	0.0	1.0	1.0	100
0.4	0.9	0.100	0.127	0.099	0.091	1.010	1.419	0.7	0.400	0.272	0.099	0.082	4.042	3.310	0.2	0.2	0.8	0.9	100
0.4	0.9	0.400	0.289	0.100	0.092	4.017	3.142	0.7	-0.400	-0.249	0.100	0.083	-4.017	-2.985	0.2	0.1	0.8	0.9	100
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

0.4	0.9	0.400	0.379	0.082	0.078	4.860	4.050	0.7	0.400	0.286	0.082	0.071	4.860	4.041	0.4	0.4	0.6	0.6	100
0.4	0.9	0.700	0.539	0.084	0.081	8.340	6.636	0.7	-0.400	-0.235	0.084	0.074	-4.766	-3.107	0.4	0.3	0.6	0.7	100
0.4	0.9	0.700	0.584	0.079	0.075	8.040	7.777	0.7	0.000	0.032	0.079	0.068	0.000	0.477	0.5	0.4	0.5	0.6	100
0.4	0.9	0.700	0.629	0.039	0.050	17.001	12.655	0.7	0.400	0.300	0.039	0.045	10.172	6.640	0.9	0.7	0.1	0.3	100
0.4	0.9	0.100	0.067	0.103	0.094	0.972	0.720	0.9	-0.400	-0.348	0.103	0.096	-3.889	-3.613	0.1	0.1	0.9	0.9	100
0.4	0.9	0.100	0.083	0.110	0.099	0.907	0.833	0.9	0.000	[illegible]	0.110	0.102	0.000	0.060	0.0	0.0	1.0	1.0	100
0.4	0.9	0.100	0.098	0.099	0.090	1.010	1.090	0.9	0.400	0.361	0.099	0.093	4.042	3.883	0.2	0.2	0.8	0.8	100
0.4	0.9	0.400	0.316	0.100	0.092	4.017	3.442	0.9	-0.400	-0.330	0.100	0.095	-4.017	-3.489	0.2	0.2	0.8	0.9	100
0.4	0.9	0.400	0.332	0.102	0.093	3.940	3.572	0.9	0.000	0.025	0.102	0.096	0.000	0.258	0.2	0.1	0.9	0.9	100
0.4	0.9	0.400	0.347	0.082	0.077	4.860	4.510	0.9	0.400	0.379	0.082	0.079	4.860	4.786	0.4	0.4	0.6	0.6	100
0.4	0.9	0.700	0.565	0.084	0.081	8.340	6.771	0.9	-0.400	-0.311	0.084	0.083	-4.766	-3.734	0.4	0.3	0.6	0.7	100
0.4	0.9	0.700	0.580	0.079	0.076	8.840	7.615	0.9	0.000	0.043	0.079	0.078	0.000	0.549	0.5	0.4	0.5	0.6	100
0.4	0.9	0.700	0.596	0.039	0.047	17.001	12.793	0.9	0.400	0.397	0.039	0.040	10.172	8.294	0.9	0.8	0.1	0.2	100

NOTATION KEY:

RT12 - CORRELATION BETWEEN TRUE X1 AND TRUE X2; R11 - RELIABILITY OF OBSERVED X1; TB1 - TRUE REGRESSION OF X1 ON Y; OB1 - OBSERVED REGRESSION OF X1 ON Y; STD1 - STANDARD ERROR FOR TRUE B1; SOB1 - STANDARD ERROR FOR OBSERVED B1; T-TB1 - T VALUE FOR TRUE B1; T-OB1 - T VALUE FOR OBSERVED B1; R22 - RELIABILITY OF TOBSERVED X2; TB2 - TRUE REGRESSION OF X2 ON Y; OB2 - OBSERVED REGRESSION OF X2 ON Y; STB2 - STANDARD ERROR FOR TRUE B2; SOB2 - STANDARD ERROR FOR OSERVED B2; T-TB2 - T VALUE FOR TRUE B2; T-OB2 - T VALUE FOR OBSERVED B2; R2T - SQUARED MULTIPLE CORRELATION FOR TRUE SCORES; R2O - SQUARED MULTIPLE CORRELATION FOR OBSERVED SCORES; VET - MEAN SQUARE ERROR FOR TRUE REGRESSION; VEO - MEAN SQUARE ERROR FOR OBSERVED REGRESSION; N - SAMPLE SIZE

NOTE: FOR ALL PARAMETERS BASED ON REPEATED DRAWINGS OF SAMPLES, I.E., STB1,SOB1,T-TB1,T-OB1,STB2,SOB2,T-TB2,T-OB2, VET, AND VEO, A CONSTANT N OF 100 WAS ASSUMED. DFE = 97 WAS USED IN ALL CALCULATIONS.

A FORTRAN COMPUTER PROGRAM FOR PERFORMING THESE KINDS OF CALUATIONS IS AVAILABLE FROM DR. NOEL DUNIVANT, ASSISTANT PROFESSOR, PSYCHOLOGY DEPARTMENT, NEW YORK UNIVERSITY, 6 WASHINGTON PLACE, 7TH FLOOR, NEW YORK, NEW YORK 10003.

APPENDIX TO CHAPTER VI

```
$JOB          G89K1309,LINES=66,NOLIST,NOWARN,NOEXT
      DIMENSION STX(2,2),SUU(2,2),TB(2),SVU(2)
C     INITIALIZE FIXED PARAMETERS
      DO 5 I=1,2
      STX(I,I)=1.0
    5 SVU(I)=0.0
      STY=1.0
      SVV=0.0
      SUU(1,2)=0.0
      SUU(2,1)=0.0
      N=100
C     READ INPUT TRUE PARAMETERS AND OBSERVED RELIABILITIES
      READ(5,101) S12LO,S12HI,S12N,R11LO,R11HI,R11N,R22LO,R22HI,
     ,R22N,B1LO,B1HI,B1N,B2LO,B2HI,B2N
  101 FORMAT(15F5.0)
      WRITE(6,102)
  102 FORMAT(1H1)
      S12INC=(S12HI-S12LO)/(S12N-1)
      STX(1,2)=S12LO-S12INC
      R11INC=(R11HI-R11LO)/(R11N-1)
      R110=R11LO-R11INC
      R22INC=(R22HI-R22LO)/(R22N-1)
      R220=R22LO-R22INC
      B1INC=(B1HI-B1LO)/(B1N-1)
      TB10=B1LO-B1INC
      B2INC=(B2HI-B2LO)/(B2N-1)
      TB20=B2LO-B2INC
      NS12=S12N
      NR11=R11N
      NR22=R22N
      NB1=B1N
      NB2=B2N
      DO 20 I=1,NS12
      STX(1,2)=STX(1,2)+S12INC
      STX(2,1)=STX(1,2)
      DO 20 J=1,NR11
      IF(J.EQ.1)R11=R110
      R11=R11 + R11INC
      SUU(1,1)=(STX(1,1)-R11*STX(1,1))/R11
      DO 20 K=1,NR22
      IF(K.EQ.1)R22=R220
      R22=R22+R22INC
      SUU(2,2)=(STX(2,2)-R22*STX(2,2))/R22
      DO 20 L=1,NB1
      IF(L.EQ.1)TB(1)=TB10
      TB(1)=TB(1)+B1INC
      DO 20 M=1,NB2
      IF(M.EQ.1)TB(2)=TB20
      TB(2)=TB(2)+B2INC
      CALL REGRES(STX,SUU,TB,STY,SVV,SVU,R11,R22,N)
   20 CONTINUE
      STOP
      END
      SUBROUTINE REGRES(STX,SUU,TB,STY,SVV,SVU,R11,R22,N)
      DIMENSION STX(2,2),SUU(2,2)TB(2),SVU(2),STXINV(2,2),STB(2,2)
     ,,TT(2),SOX(2,2),STYX(2),SOXINV(2,2),SOYX(2),SOB(2,2),
     ,OB(2),OT(2),QSTB(2),QSDB(2)
      ZN=N
C     COMPUTE TURE-SCORE PARAMETERS
C     COMPUTE R SQUARE TRUE
```

```
      CALL QUADF(TB,STX,RSQ)
      R2TY=RSQ/STY
C     COMPUTE MSE TRUE
      DF=ZN-3
      STE=STY*(ZN/DF)*(1-R2TY)
C     COMPUTE TRUE B VECTOR
      CALL MINV(STX,STXINV)
      DO 5 I=1,2
      DO 5 J=1,2
    5 STB(I,J)=STE*(1/ZN)*STXINV(I,J)
C     COMPUTE TRUE T-TESTS
      DO 10 I=1,2
      QSTB(I)=SQRT(STB(I,I))
   10 TT(I)=TB(I)/QSTB(I)
C     COMPUTE STYX - TRUE YX COVARIANCE VECTOR
      CALL MVMAT(STX,TB,STYX)
C     COMPUTE OBSERVED-SCORE PARAMETERS
C     COMPUTE SOX - SIGMA OF OBSERVED SCORES
      DO 11 I=1,2
      DO 11 J=1,2
   11 SOX(I,J)=STX(I,J)+SUU(I,J)
C     COMPUTE SOXINV
      CALL MINV(SOX,SOXINV)
C     COMPUTE OBSERVED YX COVARIANCE VECTOR
      DO 15 I=1,2
   15 SOYX(I)=STYX(I)+SVU(I)
C     COMPUTE OBSERVED-SCORE REGRESSION VECTOR
      CALL MVMAT(SOXINV,SOYX,OB)
C     COMPUTE OBSERVED Y VARIANCE
      SOY=STY + SVV
C     COMPUTE OBSERVED R SQUARE
      CALL QUADF(OB,SOX,RSQ)
      R2OY=RSQ/SOY
C     COMPUTE SOE - MSE FOR OBSERVED SCORES
      SOE=SOY*(ZN/DF)*(1-R2OY)
C     COMPUTE SIGMA OF OBSERVED B VECTOR
      DO 20 I=1,2
      DO 20 J=1,2
   20 SOB(I,J)=SOE*(1/ZN)*SOXINV(I,J)
C     COMPUTE OBSERVED T-TESTS
      DO 25 I=1,2
      QSOB(I)=SQRT(SOB(I,I))
   25 OT(I)=OB(I)/QSOB(I)
      WRITE(6,100)STX(1,2),R11,TB(1),OB(1),QSTB(1),QSOB(1),TT(1),OT(1),
     ,R22,TB(2),OB(2),QSTB(2),CSOB(2),TT(2),OT(2),R2TY,R2OY,
     ,STE,SOE,N
  100 FORMAT(1X,1F4.1,1X,1F3.1,1X,4(1F7.3,1X),2(1F8.3,1X),1F3.1,
     ,4(1F7.3,1X),2(1F8.3,1X),4(1F3.1,1X),1I3)
      RETURN
      STOP
      END
      SUBROUTINE MVMAT(X,B,XB)
C     MATRIX-VECTOR PRODUCT
      DIMENSION X(2,2),B(2),XB(2)
      DO 5 I=1,2
      XB(I)=0.0
      DO 5 J=1,2
    5 XB(I)=XB(I)+B(J)*X(J,I)
      RETURN
      STOP
      END
```

```
      SUBROUTINE MINV(X,XINV)
C     INVERT A 2X2 MATRIX
      DIMENSION X(2,2),XINV(2,2)
      D=X(1,1)*X(2,2)-X(1,2)*X(2,1)
      XINV(1,1)=X(2,2)/D
      XINV(2,2)=X(1,1)/D
      XINV(1,2)=-X(1,2)/D
      XINV(2,1)=XINV(1,2)
      RETURN
      STOP
      END
      SUBROUTINE QUADF(B,S,RSQ)
C     COMPUTE A QUADRATIC FORM: RSQ =  B' * S * B
      DIMENSION B(2),S(2,2),Y(2)
      DO 5 J=1,2
      Y(J)=0.0
      DO 5 K=1,2
    5 Y(J)=Y(J) + B(K)*S(K,J)
      DO 6 J=1,2
    6 RSQ=Y(J)*B(J)
      RETURN
      STOP
      END
$ENTRY
  -.4   .4    2  .65  .85    2   .7   .9    2   .1   .9    3  -.5   .5    3
$STOP
/*
//
```

REFERENCES

Adcock, R. J. A problem in least squares. Analyst, 1878, 5, 53-54.

Aigner, D. J. Regression with a binary independent variable subject to errors of observation. Journal of Econometrics, 1973, 1, 49-60.

Aigner, D., & Goldberger, A. S. (Eds.) Latent variables in socio-economic models. Amsterdam: North-Holland Publishing Co., 1976.

American Psychological Association. Manual for educational and psychological tests and manuals. Washington, D.C.: American Psychological Association, 1966.

Asher, H. B. Some consequences of measurement error in survey data. American Journal of Political Science, 1974, 18, 469-485.

Baltes, P. B., & Nesselroade, J. R. Cultural change and adolescent personality: An application of longitudinal sequences. Developmental Psychology, 1972, 7, 244-256.

Bartlett, M. S. On the theory of statistical regression. Proceedings of the Royal Society of Edinburgh, 1933, 53, 260-283.

Bereiter, C. Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Bergman, L. R. Some univariate models in studying change (Supplement 10). Stockholm: Psychological Laboratories, University of Stockholm, 1971.

Blalock, H. M. Path coefficients versus regression coefficients. American Journal of Sociology, 1967, 72, 675-676.

Blalock, H.M. Multiple indicators and the causal approach to measurement error. American Journal of Sociology, 1969, 75, 264-272.

Block, J. The equivalence of measures and the correction for attenuation. Psychological Bulletin, 1963, 60, 152-156.

Bock, R. D., & Petersen, A. C. A multivariate correction for attenuation. Biometrika, 1975, 62, 152-156.

Bohrnstedt, G. W. Observations on the measurement of change. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), Sociological methodology-1969. San Francisco: Jossey-Bass, 1969.

Bohrnstedt, G. W. & Carter, T. M. Robustness in regression analysis. In H. L. Costner (Ed.), Sociological methodology-1971. San Francisco: Jossey-Bass, 1969.

Box, G. E. P., & Muller, M. E. A note on the generaton of random normal deviates. Annals of Mathematical Statistics, 1958, 29, 610-611.

Brewer, M. B., Campbell, D. T., & Crano, W. D. Testing a single-factor model as an alternative to the misuse of partial correlations in hypothesis-testing research. Sociometry, 1970, 33, 1-11.

Browne, M. W. A comparison of factor analytic techniques. Psychometrika, 1968, 33, 267-334.

Busemeyer, J. R. Importance of measurement theory, error theory, and experimental design for testing the significance of interactions. Psychological Bulletin, 1980, 88, 237-244.

Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Campbell, D. T., & Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasiexperimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experiment. New York: Academic Press, 1975.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuch (Ed.), Compensatory education: A national debate. Vol. III. The Disadvantaged child. New York: Brunner/Mazel, 1970.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Cleary, T. A., & Linn, R. L. Error of measurement and the power of the statistical test. British Journal of Mathematical and Social Psychology, 1969, 22, 49-55.

Cochran, W. G. Errors of measurement in statistics. Technometrics, 1968, 10, 637-666.

Cochran, W. G. Some effects of errors of measurement on multiple correlation. Journal of the American Statistical Association, 1970, 65, 22-34.

Cochran, W. G. Some effects of errors of measurement on linear regression. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, I, 1972, 527-539.

Cohen, J. Statistical power analysis for the behavioral sciences (Rev. ed.). New York: Academic Press, 1977.

Cohen, J., & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.

Corder-Bolz, C. R. The evaluation of change: New evidence. Educational and Psychological Measurement, 1978, 38, 957-976.

Cronbach, L. J. Research on classrooms and schools: Formulation of questions, design, and analysis (Occasional Paper). Stanford, Cal.: Stanford University, Stanford Evaluation Consortium, School of Education, 1976.

Cronbach, L. J., & Furby, L. How we should measure "change"—or should we? Psychological Bulletin, 1970, 74, 68-80. See also Errata, ibid., 1970, 74, 218.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L.J., & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 456-473.

Cronbach, L. J., Rogosa, D. R., Floden, R E., & Price, G. G. Analysis of covariance in nonrandomized experiments: Parameters affecting bias. Unpublished manuscript, Stanford Educational Consortium, Stanford University, 1977.

Cureton, E. E. On certain estimated correlation functions and their standard errors. Journal of Experimental Education, 1936, 4, 252-264.

Davidson, T. N. Youth in transition. Vol. IV: Evolution of a strategy for longitudinal analysis of survey panel data. Ann Arbor, Mich.: University of Michigan, Survey Research Center 1972.

Dearborn, W. F., Rothney, J. W. M., & Shuttleworth, F. K. Data on the growth of public school children. Monographs of the Society for Research in Child Development, 1938, No. 14.

DeGracie, J. S. Analysis of covariance when the concomitant variable is measured with error (Doctoral dissertation, Iowa State University of Science and Technology, 1968). Dissertation Abstracts, 1969, 29, 3528B. (University Microfilms No. 69-04226)

DeGracie, J. S., & Fuller, W. A. Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. Journal of the American Statistical Association, 1972, 67, 930-937.

Draper, N., & Smith, H. Applied regression analysis. New York: Wiley, 1966.

DuBois, P. H. Multivariate correlational analysis. New York: Harper, 1957.

Dunivant, N. Review of least squares procedures for measuring change (Research Monograph). Austin: University of Texas, Research and Development Center for Teacher Education, 1975.

Dunivant, N. Statistical problems in the assessment of sex role development. Paper presented at meetings of the Eastern Psychological Associatio, 1977. (a)

Dunivant, N. Comments on Overall and Woodward's recommendations concerning measurement error and nonrandom assignment in the analysis of covariance. Unpublished manuscript, New York University, 1977. (b)

Dunivant, N. Estimating latent score regression when measurment error variances are known. Presented at meetings of the Eleventh Annual Symposium on the Interface of Statistics and Computer Science, Raleigh, N.C., March 1978. (a)

Dunivant, N. A comparison of statistical methods that correct for errors of measurement. Presented at meetings of the Eastern Educational Research Association, Williamsburg, Va., March 1978. (b)

Dunivant, N. Problems caused by measurement error in analyzing change. Presented at meetings of the American Educational Research Association, San Francisco, April 1979. (a)

Dunivant, N. Procedures for determining the equivalence of measures. Presented at meetings of the American Educational Research Association, San Francisco, April 1979. (b)

Dunivant, N. Correcting regression analysis for errors of measurement: A Monte Carlo study. Presented at meetings of the American Educational Research Association, Boston, April 1980. (a)

Dunivant, N. Bibliography on the effects of measurement error on linear statistical models. Unpublished manuscript, New York University, 1980. (b)

Dwyer, P. S. The square root method and its use in correlation and regression. Journal of the American Statistical Association, 1945, 40, 493-503.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-402.

Evans, S. H., & Anastasio, E. J. Misuse of analysis of covariance when treatment effect and covariate are confounded. Psychological Bulletin, 1968, 69, 225-234.

Fix, E. Distributions which lead to linear regressions. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1949.

Forsyth, R.A., & Feldt, L.S. An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. Educational and Psychological Measurement, 1969, 29, 61-72.

Forsyth, R.A., & Feldt, L.S. Some theoretical and empirical results related to McNemar's test that the population correlation coefficient corrected for attenuation equals 1.0. American Educational Research Journal, 1970, 7, 197-207.

Friedrich, D. D., & Van Horn, K. R. Developmental methodology: A revised primer. Minneapolis, Minn.: Burgess, 1976.

Fuller, W.A. Properties of some estimators for the errors-in-variables model. Paper presented at meetings of the Econometric Society, New Orleans, December 1971.

Fuller, W.A. Regression analysis for sample survey. Sankhya, 1975, 37C, 117-132.

Fuller, W. A. Properties of some estimators for the errors in variables model. Annals of Statistics, 1980, in press.

Fuller, W. A. Estimation of measurement error models from cluster samples. Presented at meetings of the American Educational Research Association, Boston, April 1980.

Fuller, W. A., & Booth, G. D. The errors-in-variables model when the covariance matrix is not constant. Proceedings of the American Statistical Association, Business and Economics Statistics Section, 1972.

Fuller, W. A., & Hidiroglou, M. A. Properties of the correction for attenuation estimator. Proceedings of the American Statistical Association, Social Statistics Section, Part I, 1976, 340-343.

Fuller, W. A., & Hidiroglou, M. A. Regression estimation after correcting for attenuation. Journal of the American Statistical Association, 1978, 73, 99-104.

Games, P. A. Confounding problems in multifactor AOV when using several organismic variables of limited reliatiblity. American Education Research Journal, 1975, 12, 225-232.

Goldberger, A. S. Structural equation models: An overview. In A. S. Goldberger & O. D. Duncan (Eds.), Structural equation models in the social sciences. New York: Seminar Press, 1973.

Goldberger, A.S., & Duncan, O.D. Structural equation models in the social sciences. New York: Seminar Press, 1973.

Graybill, F. A. An introduction to linear statistical models (Vol. 1). New York: McGraw-Hill, 1961.

Hamilton, B. L. A Monte Carlo test of the robustness of parametric and nonparametric analysis of covariance against unequal regression slopes. Journal of the American Statistical Association, 1976, 71, 864-869.

Hammersley, J.M., & Handscomb, D.C. Monte Carlo methods. Methuen, 1964.

Hanushek, E. A., & Jackson, J. E. Statistical methods for social scientists. New York: Academic Press, 1977.

Harnqvist, K. Correcting analysis of covariance for unreliability of the control variable. Unpublished report, University of Gothenburg, Sweden, 1958.

Heckman, J. Sample selection as a specification error. Econometrica, 1979, 47, 153-161.

Hidiroglou, M. A. Estimation of regression parameters for finite samples. Unpublished doctoral dissertation, Iowa State University of Science and Technology, 1976.

Hidiroglou, M. A., Fuller, W. A., & Hickman, R. D. SUPER CARP (Cluster Analysis and Regression Program). Unpublished report, Statistical Laboratory, Iowa State University, 1977.

Hilton, T. L. Growth study annotated bibliography (Progress Report 69-11). Princeton, N.J.: Educational Testing Service, 1969.

Hotelling, H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 1933, 24, 417-441, 498-520.

Hsu, T.C., & Sebatane, E.M. The effect of differences in covariate means among treatment groups upon the F-test of the analysis of covariance. Paper presented at the Annual Meeting of the American Educational Research Association, New York, N.Y., April, 1977.

Hummel-Rossi, B., & Weinberg, S. L. Practical guidelines in applying current theories to the measurement of change. Part I: Problems in measuring change and recommendedprocedures. Part II: Numerical illustrations of research designs concerned with measuring change. JSAS Catalog of Selected Documents in Psychology, 1975, 5, 226. (Ms No. 916)

Hunter, J. E., & Cohen, S. H. Correcting for unreliability in nonlinear models of attitude change. Psychometrika, 1974, 39, 445-468.

International Mathematical & Statistical Libraries, Inc. (IMSL). The IMSL library (3 vols.). Houston, Texas: Author, 1980.

Johnston, J. Econometric methods. New York: McGraw-Hill, 1963.

Johnston, J. Econometric methods (2nd ed.). New York: McGraw-Hill, 1972.

Jöreskog, K. G. A general method for analysis of covariance structures. Biometrika, 1970, 57, 239-251.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36, 109-133.

Jöreskog, K. G. A general method for estimating a linear structural equation system. In A.S. Goldberger & O.D. Duncan (Eds.), Structural equation models in the social sciences. New York: Seminar Press, 1973.

Jöreskog, K. G. Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology (Vol. II). San Francisco: Freeman, 1974.

Jöreskog, K. G., & Sörborn, D. Some regression estimates useful in the measurement of change. Research report 74-2. Uppsala: Uppsala University, Statistics Department, 1974.

Jöreskog, K. G., & Sörborn, D. Statistical models and methods for test-retest situations. In D. N. M. de Gruijter & L. J. Th. van der Kamp (Eds.), Advances in psychological and educational measurement. London: Wiley, 1976.

Jöreskog, K. G., & Sörborn, D. Statistical models and methods for analysis of longitudinal data. In D. J. Aigner & A. S. Goldberger (Eds.), Latent variables in socio-economic models. Amsterdam: North Holland Publishing Co., 1977.

Jöreskog, K. G., & Sörborn, D. LISREL - A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables (ETS RB 72-56). Princeton, N.J.: Educational Testing Service, 1972.

Kagan, J. American longitudinal research in psychological development. Child Development, 1964, 35, 1-32.

Kagan, J., & Moss, H. A. Birth to Maturity: The Fels study of psychological development. New York: Wiley, 1962.

Kahneman, D. Control of spurious association and the reliability of the controlled variable. Psychological Bulletin, 1965, 64, 326-329.

Keesling, J. W., & Wiley, D. Measurement error and the analysis of quasi-experimental data (Report No. 11). Chicago: ML-Group for Policy Studies in Education, CEMEREL, Inc., 1977.

Kelley, T.L. Fundamentals of statistics. Cambridge: Harvard University Press, 1947.

Kendall, M. G. Regression, structure and functional relationship: Part I. Biometrika, 1951, 38, 11-25.

Kendall, M. G. Regression, structure and functional relationship: Part II. Biometrika, 1952, 39, 96-108.

Kendall, M. G., & Stuart, A. The advanced theory of statistics. Vol. II: Inference and relationship. London: Charles Griffin and Co., Ltd., 1961.

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control groups design. Psychological Bulletin, 1975, 82, 345-362.

Koopmans, T.C. Linear regression analysis in economic time series. Haarlem: De erven F. Bohn N.V., 1937.

Kshirsagar, A. M. Bartlett decomposition and Wishart distribution. Annals of Mathematical Statistics, 1959, 30, 239-241.

Kummel, C. H. Reduction of observation equations which contain more than one observed quantity. Analyst, 1879, 6, 97-105.

Ladd, G. W. Effects of shocks and errors in estimation: An empirical comparison. Journal of Farm Economics, 1956, 38, 485-495.

Lawley, D. N., & Maxwell, A. E. Regression and factor analysis. Biometrika, 1973, 60, 331-338.

Lindley, D. V. Regression lines and the linear functional relationship. Journal of the Royal Statistical Society (Supplement), 1947, 9, 218-244.

Linn, R. L., & Slinde, J. A. The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 1977, 47, 121-150.

Linn, R. L., & Werts, C. E. Errors of inference due to errors of measurement. Educational and Psychological Measurement, 1973, 33, 531-543.

Lord, F. M. The measurement of growth. Educational and Psychological Measurement, 1956, 16, 421-437. See also Errata, ibid., 1957, 17, 452.

Lord, F.M. A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 1957, 22, 207-220.

Lord, F.M. Significance test for a partial correlation corrected for attenuation. Educational and Psychological Measurement, 1974, 34, 211-220.

Lord, F. M. Further problems in the measurement of growth. Educational and Psychological Measurement, 1958, 18, 437-454.

Lord, F. M. Large-sample covariance analysis when the control variable is fallible. Journal of the American Statistical Association, 1960, 55, 309-321.

Lord, F. M. Automated hypothesis and standard errors for nonstandard problems. The American Statistician, 1975, 29, 56-59.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison, WI: University of Wisconsin Press, 1963.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores (with contributions by Allan Birnbaum). Reading, Mass: Addison-Wesley, 1968.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

Madansky, A. The fitting of straight lines when both variables are subject to error. Journal of the American Statistical Association, 1959, 54, 173-205.

Marston, P. T., & Borich, G. D. Analysis of covariance: Is it the appropriate model to study change? Paper presented at the Annual Meeting of the American Educational Research Association, New York City, April, 1977.

McCallum, B. T. Relative asymptotic bias from errors of omission and measurement. Econometrica, 1972, 40, 757-758.

McLean, J. E., Ware, W. B., & McClave, J. T. How analysis of covariance can yield misleading results in educational experiments--A Monte Carlo study. Proceedings of the American Statistical Association, Social Statistics Section, 1975, 548-551.

McNemar, Q. Attenuation and interaction. Psychometrika, 1958, 23, 259-265.

McNemar, Q. On growth measurement. Educational and Psychological Measurement, 1958, 18, 47-55.

Mendenhall, W., & Scheaffer, R.L. Mathematical statistics with applications. North Scituate, Mass.: Duxbury Press, 1973.

Meredith, W. Canonical correlations with fallible data. Psychometrika, 1964, 29, 55-65.

Mouchart, M. A regression model with an explanatory variable which is both binary and subject to errors (CORE Discussion Paper 7227). Louvain: Center for Operations Research and Econometrics, 1972.

Mouchart, M. A regression model with an explanatory variable which is both binary and subject to erros. In D. J. Aigner & A. S. Goldberger (Eds.), Latent variables in socio-economic models. Amsterdam: North-Holland, 1977.

Mulaik, S. A. The foundations of factor analysis. New York: McGraw Hill, 1972.

Muller, M. E. A comparison of methods for generating normal deviates on digital computers. Journal of the Association for Computing Machinery, 1959, 6, 376-383.

Murray, J. R. Statistical models for qualitative data with classification errors. Doctoral dissertation, University of Chicago, Committee on Human Development, 1971.

Namboodiri, N. K., Carter, L. F., & Blalock, H. M., Jr. Applied multivariate analysis and experimental designs. New York: McGraw-Hill, 1975.

Nicewander, A., & Price, J. M. Dependent variable reliability and tests of significance. Paper presented at the Annual Meeting of the Psychometrika Society, Chapel Hill, N.C., June, 1977.

O'Connor, E. F., Jr. Extending classical test theory to the measurement of change. Review of Educational Research, 1972, 42, 73-97.

Odell, P. L., & Feiveson, A. H. A numerical procedure to generate a sample covariance matrix. Jouranl of the American Statistical Association, 1966, 61, 199-203.

Olejnik, S. F., & Porter, A. C. Bias and mean square errors of estimators as criteria for evaluating competing analysis strategies in quasi-experiments. Journal of Educational Statistics, 1981, 6, 33-53.

Overall, J. E., & Klett, C. J. Applied multivariate analysis. New York: McGraw-Hill, 1972.

Overall, J. E., & Woodward, J. A. Nonrandom assignment and the analysis of covariance. Psychological Bulletin, 1977, 84, 588-594. (a)

Overall, J. E., & Woodward, J. A. Common misconceptions concerning the analysis of covariance. Multivariate Behavioral Research, 1977, 12, 171-185. (b)

Porter, A. C. The effects of using fallible variables in the analysis of covariance (Doctoral dissertation, University of Wisconsin, 1967). Dissertation Abstracts, 1967, 28, 3517B. (University Microfilms No. 67-12147)

Porter, A. C., & Chibucos, T. R. Selecting analysis strategies. In G. D. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Pearson, K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 1901, 2, 559-572.

Peckham, P. D. The robustness of the analysis of covariance to heterogeneous regression slopes. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, March, 1970.

Richards, J. M., Jr. A simulation study of the use of change measures to compare educational programs. American Education Research Journal, 1975, 12, 299-311.

Rindskopf, D. M. A comparison of various regression-correlation methods for evaluating nonexperimental research (Doctoral dissertation, Iowa State University of Science and Technology, 1976). Dissertation Abstracts International, 1977, 37, 4120B. (University Microfilms No. 77-01, 471)

Rogosa, D. Measurement error and the Johnson-Neyman technique. Paper presented at the Annual Meetings of the American Educational Research Association, New York, April, 1977. (a)

Rogosa, D. Some results for the Johnson-Neyman technique. Doctoral dissertation, Standford University, School of Education, 1977.

Rozeboom, W. W. Dynamic analysis of multivariate process data. Unpublished manuscript, University of Alberta, Psychology Department, January 1977.

Rubin, D. B. Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics, 1977, 2, 1-26.

Saunders, D. R. Partial reliability coefficients applied to the overcorrection problem (RM 51-16). Princeton, N.J.: Educational Testing Service, 1951.

Schmidt, P. Econometrics. New York: Marcel Dekker, 1976.

Schreider, Yu. A. The Monte Carlo method: The method of statistical trials. Oxford: Pergamon, 1966. (Translated from the Russian by G. J. Tee.)

Shen, E. The standard error of certain estimated coefficients of correlation. Journal of Educational Psychology, 1924, 15, 462-465.

Snedecor, G. W., & Cochran, W. G. Statistical methods (6th ed.). Ames, Iowa: University of Iowa Press, 1967.

Sörbom, D. An alternative to the methodology for analysis of covariance. Psychometrika, 1978, 43, 381-396.

Spearman, C. The proof and measurement of association between two things. American Journal of Psychology, 1904, 15, 72-101. (a)

Stebbins, L. B., Bock, G., & Proper, E. C. Educational as experimentation: A planned variation model (Vol. IV-B). Cambridge, Mass.: Abt Associates, Inc., 1977.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B., & Cerva, T. R. Education as experimentation: A planned variation model (Vol. IV-A). Cambridge, Mass.: Abt Associates, Inc., 1977.

Stocking, M., & Lord, F. M. AUTEST—Program to perform automated hypothesis tests for nonstandard problems (RM-74-25). Princeton: Educational Testing Service, 1974.

Stouffer, S. A. Evaluating the effect of inadequately measured variables in partial correlation analysis. Journal of the American Statistical Association, 1936, 31, 348-360. (a)

Stouffer, S. A. Reliability coefficients in a correlation matrix. Psychometrika, 1936, 1, 17-20. (b)

St. Pierre, R. G., & Ladner, R. Correcting covariates for unreliability: Does it lead to differences in an evaluator's conclusions? Paper presented at the meeting of the American Educational Research Association, New York, April 1977.

Stroud, T. W. F. Comparing conditional distributions under measurement errors of known variances (Unpublished doctoral dissertation, Stanford University, 1968). Dissertation Abstracts International, 1969, 29, 2673B. (University Microfilms No. 69-00299)

Stroud, T. W. F. Comparing conditional means and variances in a regression model with measurement errors of known variances. Journal of the American Statistical Association, 1972, 67, 407-412.

Stroud, T. W. F. Comparing regressions when measurement error variances are known. Psychometrika, 1974, 39, 53-68.

Sutcliffe, J. P. Error of measurement and the sensitivity of a test of significance. Psychometrika, 1958, 23, 9-17.

Theil, H. Economic forecasts and policy. Amsterdam: North-Holland, 1958.

Theil, H. Principles of econometrics. New York: Wiley, 1971.

Thomson, G. H. A formula to correct for the effect of errors of measurement on the correlation of initial values with gains. Journal of Experimental Psychology, 1924, 7, 321-324.

Thorndike, R. L. Regression fallacies in the matched groups experiment. Psychometrika, 1942, 7, 85-102.

Tucker, L. R., Damarin, F., & Messick, S. A basefree measure of change. Psychometrika, 1966, 31, 457-473.

Van de Geer, J. P. Introduction to multivariate analysis for the social sciences. San Francisco: Freeman, 1971.

Veldman, D. J. Empirical tests of statistical assumptions (RMM-9R). Austin, Texas: University of Texas, Research and Development Center for Teacher Education, 1972.

Wald, A. Test of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 1943, 54, 426-482.

Walker, H. M., & Lev, J. Statistical inference. New York: Henry Holt & Co., 1953.

Warren, R.D., White, J.K., & Fuller, W.A. An errors-in-variables analysis of managerial-role performance. Journal of the American Statistical Association, 1974, 69, 886-893.

Weisberg, H. I. Statistical adjustments and uncontrolled studies. Psychological Bulletin, 1979, 86, 1149-1164.

Werts, C. E., & Hilton, T. L. Intellectual status and intellectual growth, again. American Educational Research Journal, 1977, 14, 137-146.

Werts, C. E., & Linn, R. L. A general linear model for studying growth. Psychological Bulletin, 1970, 73, 17-22.

Werts, C. E., & Linn, R. L. Path analysis: Psychological examples. Psychological Bulletin, 1970, 74, 193-212.

Werts, C. E., & Linn, R. L. Analyzing school effects: ANCOVA with a fallible covariate. Educational and Psychological Measurement, 1971, 31, 95-104.

Werts, C.E., Linn, R.L., & Jöreskog, K.G. Another perspective on "Linear regression, structural relations, AND measurement error." Educational and Psychological Measurement, 1973, 33, 32 -332.

Wijsman, R. A. Applications of a certain representation of the Wishart matrix. Annals of Mathematical Statistics, 1959, 30, 597-601.

Wiley, D. E., & Harnischfeger, A. Post hoc, ergo propter hoc: Problems in the attribution of change (Report No. 5). Chicago, Ill.: University of Chicago, Studies of Educative Processes, 1973.

Wiley, D. E., & Hornik, R. Measurement error and the analysis of panel data (Report No. 7). Chicago, Ill.: University of Chicago, Studies of Educative Processes, 1973.

Wiley, D. E., Schmidt, W. H., & Bramble, W. J. Studies of a class of covariance structure models. Journal of the American Statistical Association, 1973, 58, 317-323.

Witkin, H. A., Goodenough, D. R., & Karp, S. A. Stability of cognitive style from childhood to young adulthood. Journal of Personality and Social Psychology, 1967, 7, 291-300.

Wolter, K. M. Estimators for a nonlinear functional relationship. Unpublished doctoral dissertation, Iowa State University, 1974.

Wolter, K. M., & Corby, C. New software for the linear errors-in-variables model. Proceedings of the American Statistical Association, 1976.

Wolter, K. M., & Fuller, W. A. Estimation of the quadratic errors-in-variables model. Unpublished report, U.S. Bureau of the Census, Washington, D. C., 1977.

Wolter, K. M., & Fuller, W. A. Estimation of the nonlinear errors-in-variables model. Unpublished report, U.S. Bureau of the Census, Washington, D. C., 1977.